BIG DATA

A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK

大数据时代

生活、工作与思维的大变革

[英]维克托·迈尔-舍恩伯格（Viktor Mayer-Schönberger） 肯尼思·库克耶（Kenneth Cukier）◎著 盛杨燕 周涛◎译

浙江人民出版社
ZHEJIANG PEOPLE'S PUBLISHING HOUSE

BIG DATA

A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE,WORK,AND THINK

大数据时代

生活、工作与思维的大变革

[英] 维克托·迈尔-舍恩伯格（Viktor Mayer-Schönberger）
肯尼思·库克耶（Kenneth Cukier）◎著
盛杨燕 周涛◎译

浙江人民出版社
ZHEJIANG PEOPLE'S PUBLISHING HOUSE

版权信息

本书纸版由浙江人民出版社于2012年12月出版

作者授权湛庐文化（Cheers Publishing）作中国大陆（地区）电子版发行（限简体中文）

书名：大数据时代：生活、工作与思维的大变革

著者：（英）维克托·迈尔-舍恩伯格,肯尼思·库克耶

字数：215000

电子书定价：24.99美元



《经济学人》说，在大数据领域，他是最受人尊敬的权威发言人之一；

《科学》说，若要发起一场关于这个问题的深入讨论，没有比他更好的发起者了。

他是欧盟互联网官方政策背后的重要制定者与参与者；

他是最早洞见大数据时代发展趋势的数据科学家之一；

他，就是 **维克托·迈尔－舍恩伯格**。

孜孜不倦的数据科学家

维克托·迈尔－舍恩伯格二十多年来一直致力于网络经济、信息与创新、信息监管、网络规范与战略管理的研究。
从维也纳大学到哈佛大学，从新加坡国立大学到牛津大学，世界上最著名的互联网研究学府都留下了他的足迹。

他是哈佛大学肯尼迪学院信息监管科研项目负责人，哈佛国家电子商务研究中心网络监管项目负责人。在哈佛大学任教10年后，2008年，他远渡重洋，来到新加坡国立大学担任信息政策研究中心主任，以崭新的视角洞察亚洲信息政策。期间还担任耶鲁大学、弗吉尼亚大学、圣地亚哥大学、芝加哥大学、维也纳大学等多所知名学府的客座教授。

而现在，在数据信息界孜孜不倦求索的维克托，是世界著名学府牛津大学网络学院互联网研究所治理与监管专业教授，牛津大学克伯学院教授研究员。法学、信息学与网络等多学科的跨界融合，让他不仅拥有严谨的思维，更拥有广博的视野。他先后有100多篇论文公开发表在《科学》《自然》等著名学术期刊上，同时也是多家出版机构的特约评论员，包括哈佛大学出版社、麻省理工学院出版社、通信政策期刊、美国社会学期刊等。

开大数据系统研究之先河

他说，世界的本质就是数据，大数据将开启一次重大的时代转型；
他说，大数据发展的核心动力来源于人类测量、记录和分析世界的渴望。
他说，从因果关系到相关关系的思维变革才是大数据的关键，建立在相关关系分析法基础上的预测才是大数据的核心。

当大数据还在众说纷纭时，维克托早已进行了系统而深入的研究。早在2010年，他已经在《经济学人》上和数据编辑肯尼思•库克耶一起，发表了长达14页的大数据专题文章，成为最早洞见大数据时代发展趋势的数据科学家之一。

从大数据时代的思维变革到商业变革，从大数据时代的隐忧到管理变革，他在大数据的蓝色海洋中预见了正在发生的未来。

大数据取舍之道

维克托的继父去世时留下了16 000张收藏的照片，这些照片都是他几十年来周游世界时的影像记录。

为了确定哪些照片需要保留，维克托制定了两条规则：首先，照片上有认识或可能认识的人；其次，照片拍摄得很漂亮。最终，他只留下了53张照片……

也许正是这一人生的片段，让他开始重新思考“被遗忘的权利”与互联时代的取舍之道，并最终成就了他的《删除》一书：“过去正像刺青一样被刻在我们的数字皮肤上，遗忘已经变成了例外，而记忆却成了常态……”大数据时代，人类又该如何构建积极而安全的未来？

《删除》一经出版，即获得美国政治科学协会颁发的“唐•K•普赖斯奖”以及媒介环境学会的最高荣誉“马歇尔•麦克卢汉奖”，同时受到《连线》《自然》《华尔街日报》《纽约时报》等各大权威媒体的广泛好评。

大数据商业应用的引路人

维克托是真正的实战派，早在上大学期间，他就先后创立了两家数据安全和制作反病毒软件的公司，并担任总裁兼CEO。

由他的公司开发的反病毒程序，一经推出即一跃成为奥地利当时最畅销的软件产品。1991年，他跻身奥地利软件企业家前5名之列，2000年成为奥地利萨尔斯堡州的年度人物。

开阔的学术视野与系统的学术造诣，更让他不断为企业与商业应用提供强大的理论支持。他的咨询客户包括微软、惠普、IBM、亚马逊、facebook、twitter、VISA等大数据先锋们，所以在《大数据时代》一书中，他才能掌握最前沿、最崭新的大数

据应用案例，并对大数据的价值链与角色定位给予清晰的预见。

欧盟专家与真正的“亚洲通”

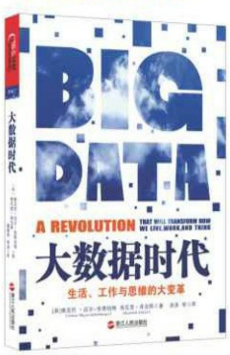
“他是欧盟互联网官方政策背后真正的制定者与参与者；他是世界经济论坛、马歇尔计划基金会等重要机构的策略顾问；他是众多国家政府高层的信息政策智囊，包括俄罗斯商务部高层、瑞士联邦政府、荷兰政府高层等。”

尤为重要的是，他特别熟悉亚洲信息产业的发展与战略布局，一直专注于信息安全、信息政策与战略的研究，先后担任新加坡商务部高层、文莱国防部高层、科威特商务部高层、迪拜及中东政府高层的咨询顾问。

正如大数据提供的不是最终答案，只是参考答案一样，大数据正在改变我们的生活以及理解世界的方式，正在成为新发明和新服务的源泉，而维克托·迈尔-舍恩伯格更多的大数据研究也在蓄势待发……

维克托·迈尔-舍恩伯格 作品系列

大数据时代



删 除



湛庐文化 Cheers Publishing
特别制作

目录

推荐序一 拥抱“大数据时代”

宽带资本董事长 田溯宁

从硅谷到北京，大数据的话题正在被传播。随着智能手机以及“可佩带”计算设备的出现，我们的行为、位置，甚至身体生理数据等每一点变化都成为了可被记录和分析的数据。以此为基础，“反馈经济”（feedback economy）等新经济、新商业模式也正在开始形成。维克托·迈尔-舍恩伯格教授这本《大数据时代》，是我看到的最好的大数据著作，不管对于产业实践者，还是对于政府和公众机构，都是非常具有价值的。

如今，一个大规模生产、分享和应用数据的时代正在开启。正如维克托教授所说，大数据的真实价值就像漂浮在海洋中的冰山，第一眼只能看到冰山的一角，绝大部分都隐藏在表面之下。而发掘数据价值、征服数据海洋的“动力”就是云计算。互联网时代，尤其是社交网络、电子商务与移动通信把人类社会带入了一个以“PB”（1024TB）为单位的结构与非结构数据信息的新时代。在云计算出现之前，传统的计算机是无法处理如此量大、并且不规则的“非结构数据”的。以云计算为基础的信息存储、分享和挖掘手段，可以便宜、有效地将这些大量、高速、多变化的终端数据存储下来，并随时进行分析与计算。大数据与云计算是一个问题的两面：一个是问题，一个是解决问题的方法。通过云计算对大数据进行分析、预测，会使得决策更为精准，释放出更多数据的隐藏价值。数据，这个21世纪人类探索的新边疆，正在被云计算发现、征服。

《大数据时代》列举了众多在公共卫生、商业服务领域大数据变革的例子。一旦“不再追求精确度，不再追求因果关系，而是承认混杂性，探索相关关系”，“思维转变过来，数据就能被巧妙地用来激发新产品和新型服务”。数据正成为巨大的经济资产，成为新世纪的矿产与石油，将带来全新的创业方向、商业模式和投资机会。

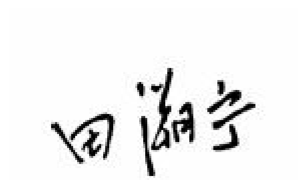
庞大的人群和应用市场，复杂性高、充满变化，使得中国成为世界上最复杂的大数据国家。解决这种由大规模数据引发的问题，探索以大数据为基础的解决方案，是中国产业升级、效率提高的重要手段。数据挖掘不仅能够成为公司竞争力的来源，也将成为国家竞争力的一部分。联系到我国现代化所面临的种种问题以及教育、交通、医疗保健等各方

面挑战，通过大数据这种创新方式来解决问题，创建新的产业群，实现“中国制造到中国创造”的改变，意义就更大。

“大数据”发展的障碍，在于数据的“流动性”和“可获取性”。美国政府创建了Data.gov网站，为大数据敞开了大门；英国、印度也有“数据公开”运动。中国要赶上这样一场大数据变革，各界应该首先开始尝试公开数据、方式与方法。如同工业革命要开放物质交易、流通一样，开放、流通的数据是时代趋势的要求。《大数据时代》一书也提到了数据拥有权、隐私性保护等问题，但相比较来看，新科技可能带来的改变要远远大于其存在的问题。

本书的译者周涛教授是我国最年轻有为的大数据专家。这位27岁的天才型教授，数年来一直带领我国学术界在大数据研究上向国际一流看齐。更可贵的是，他不仅做研究，也关注着研究成果的商业化及传播。这部译著就是他这种努力的一个成果。

现代历史上的历次技术革命，中国均是学习者。而在这次云计算与大数据的新变革中，中国与世界的距离最小，在很多领域甚至还有着创新与领先的可能。只要我们以开放的心态、创新的勇气拥抱“大数据时代”，就一定会抓住历史赋予中国创新的机会。



推荐序二 实实在在大数据

中国互联网发展的重要参与者，知名IT评论人 谢文

因为我本身十分关注大数据，也写过若干关于大数据的文章，做过若干关于大数据的演讲，所以对有关这一主题的论文和书籍非常有兴趣。过去几年，在这方面我读过十几本书、上百篇论文和文章。相对而言，维克托·迈尔-舍恩伯格教授的《大数据时代》是迄今为止我读过的最好的一本专著，中英文都算上。

此书的一大贡献就是在大数据方兴未艾、众说纷纭的时刻，进一步阐述和厘清了大数据的基本概念和特点，这对许多以为大数据就是“数据大”的人来说很有帮助。

在人类历史长河中，即使是在现代社会日新月异的发展中，人们还主要是依赖抽样数据、局部数据和片面数据，甚至在无法获得实证数据的时候纯粹依赖经验、理论、假设和价值观去发现未知领域的规律。因此，人们对世界的认识往往是表面的、肤浅的、简单的、扭曲的或者是无知的。维克托指出，大数据时代的来临使人类第一次有机会和条件，在非常多的领域和非常深入的层次获得和使用全面数据、完整数据和系统数据，深入探索现实世界的规律，获取过去不可能获取的知识，得到过去无法企及的商机。

大数据的出现，使得通过数据分析获得知识、商机和社会服务的能力从以往局限于少数象牙塔之中的学术精英圈子扩大到了普通的机构、企业和政府部门。门槛的降低直接导致了数据的容错率提高和成本的降低，但正如维克托所强调的，最重要的是人们可以在很大程度上从对于因果关系的追求中解脱出来，转而将注意力放在相关关系的发现和使用上。只要发现了两个现象之间存在的显著相关性，就可以创造巨大的经济或社会效益，而弄清二者为什么相关可以留待学者们慢慢研究。大数据之所以可能成为一个“时代”，在很大程度上是因为这是一个可以由社会各界广泛参与，八面出击，处处结果的社会运动，而不仅仅是少数专家学者的研究对象。

大数据将逐渐成为现代社会基础设施的一部分，就像公路、铁路、港口、水电和通信网络一样不可或缺。但就其价值特性而言，大数据却

和这些物理化的基础设施不同，不会因为人们的使用而折旧和贬值。例如，一组DNA可能会死亡或毁灭，但数据化的DNA却会永存。所以，维克托赞同许多物理学家的看法，世界的本质就是数据。因此，大数据时代的经济学、政治学、社会学和许多科学门类都会发生巨大甚至是本质上的变化和发展，进而影响人类的价值体系、知识体系和生活方式。哲学史上争论不休的世界可知论和不可知论将会转变为实证科学中的具体问题。可知性是绝对的，无事无物不可知；不可知性是相对的，是尚未知道的意思。

对于不从事网络业、IT业以及数据分析和使用的读者，本书的一大好处就是通俗易懂，通过具体实例说明问题，有助于人们的理解和联想。在时限上，作者概括了直到2012年7月大数据方向上的最新发展，避免了许多同类作品存在的例证过于陈旧、视野相对狭窄的毛病。

作为一位生活在欧美现代社会的学者，维克托是把民主、开放和理性作为已知前提来讨论大数据革命的。这对生活在发展中国家，社会现代化程度尚且有限的读者来说，也许是个遗憾，因为书中描述的许多已经发生的事例可能更像是神话。没有市场经济制度和法治体系作为基础支撑，大数据很可能成为发达国家在下一轮全球化竞争中的利器，而发展中国家依然处于被动依附的状态之中。整个世界可能被割裂为大数据时代、小数据时代和无数据时代。

处于发展中国家前列的中国，目前正面临着一个重大的历史抉择关口。应该说，在过去的三十余年时间里，中国在快速走向工业化、信息化、网络化方面交出了一份不错的成绩单。如今适逢世界走向数据化，迈入大数据时代的时刻，无论对个人、企业还是对社会和国家，都有认真理解、严肃决策的必要性和紧迫性。哪怕仅从这一点考虑，读一读这本书也是很值得的。

译者序
在路上·晃晃悠悠

电子科技大学教授，互联网科学中心主任 周涛

接下翻译这本《大数据时代》的任务时，我的目标是做到110%的好。因为作者维克托·迈尔-舍恩伯格毕竟不像我们每天在一线与数据厮杀搏斗，其爱其恨都更深刻。特别地，我们可以为中文的读者补充很多中国的例子和参考资料。很遗憾，我们最终只做到了90%，应该补充的一些材料还没有整理好，遣词造句也多有生硬疏忽之处。如果再给我一个月的时间，就可以达到我预想的110%甚至120%。

为什么现在把这个版本呈现给诸位呢？一是因为我们的努力使得本书中译本的出版和英文原版完全同步，单从获取知识的角度讲，我们一点儿不比美国的读者慢！二是我相信作者在书中的一个重要观点，就是大数据时代，要允许一点点的错误和不完美，因为效率可能更加重要！留下一些可供提高的地方，也使得我们的每一次印刷，都能够与以前有所不同。亲，这不是建议你等到某个更好的版本才去购买，而是说，其实你应该每个版本都买一本：）

《大数据时代》这本书是200%的好，因此90%的译本也绝对值得一读。首先，作者抛出了大数据时代处理数据理念上的三大转变：要全体不要抽样，要效率不要绝对精确，要相关不要因果；接着，从万事万物数据化和数据交叉复用的巨大价值两个方面，讲述驱动大数据战车在材质和智力方面向前滚动的最根本动力；最后，作者冷静描绘了大数据帝国前夜的脆弱和不安，包括产业生态环境、数据安全隐私、信息公正公开等问题。

国内最近也出版了一些大数据方面的著作，可以和本书互为补充。郑毅的《证析》对于数据通过交叉复用体现的新价值、大数据战略在企业与政府执行层面的流程和大数据科学家这一新职位，以及围绕这个职位的能力和责任给出了最深刻、最具体的描述；子沛的《大数据》对于数据的公正性、公平性以及信息和数据管理等方面理念、政策和执行的变化，特别是美国在这方面的进展，给出了完整的介绍；苏萌、林森和我合著的《个性化：商业的未来》则对大数据时代最重要的技术——个

性化技术，以及与之相关的新商业模式给出了从理念到技术细节的全景工笔。总的来说，这三本书都针对本书的某一局部给出了更深刻的介绍和洞见，也各有明显超出本书的优点，但三本之和也无法囊括本书的菁华，亦缺乏本书的宏大视野。

简单地说，这本书好在三个地方：

一是观点掷地有声，绝非主流媒体上若干讨论的简单汇总和平均，更不是一个宏大概念面前暧昧的叫好声。读者可能对其中一些观点并不认同，但是读完之后不可能一个都记不住。

二是观念高屋建瓴，作者试图从很多实例和经验，包括历史事件中萃取出普适性的观念，而不仅仅是适用于几个特定情况的案例分析。

三是例子丰富翔实，不长的篇幅包括了上百个学术和商业的实例。

三点近乎完美地结合起来，体现了作者驾驭大问题的能力和丰富的知识，以及，可能更为重要地，作者渴求立言立说的野心！所以说，这本书绝对不是一堆枯燥的纲要，更不是一本巨厚的杂志。

我在这里拼命叫好，是为了这本书卖得更多，但不代表作者的所有观点都是绝对真理。举个例子，我本人对于大数据时代“相关关系比因果关系更重要”这个观点就不认同。有了机器学习，特别是集成学习，我们解决问题的方式变成了训练所有可能的模型和拟合所有可能的参数——问题从一个端口进去，答案从另一个端口出来，中间则是一个黑匣子，因为没有人能够从成千上万的参数拟合值里面读到“科学”，我们读到的只是“计算机工程”。与其说大数据让我们重视相关胜于因果，不如说机器学习和以结果为导向的研究思路让我们变成这样。

那么，大数据是不是都这样呢？其实很多时候恰恰相反。想想瑞士日内瓦的强子对撞机，我们在上面捕获了人类有史以来最大规模的单位时间数据。我们是希望找到或者验证某种相关关系吗？不是！我们试图回答的，正是人类所能问出的关于因果关系最伟大的问题：希格斯玻色子是否存在，我们的宇宙是否有可能用标准模型刻画。这个问题的最终答案，将打破人和神的界限！认为相关重于因果，是某些有代表性的大数据分析手段（譬如机器学习）里面内禀的实用主义的魅影，绝非大数据自身的诉求。从小处讲，作者试图避免的“数据的独裁”和“错误的前提导致错误的结论”，其解决之道恰在于挖掘因果逻辑而非相关性；从

大处讲，放弃对因果性的追求，就是放弃了人类凌驾于计算机之上的智力优势，是人类自身的放纵和堕落。如果未来某一天机器和计算完全接管了这个世界，那么这种放弃就是末日之始。

苏珊·朗格（Susan Langer）在《哲学新视野》一书中说：

某些观念有时会以惊人的力量给知识状况带来巨大的冲击。由于这些观念能一下子解决许多问题，所以，它们似乎将有希望解决所有基本问题，澄清所有不明了的疑点。每个人都想迅速地抓住它们，作为进入某种新实证科学的法宝，作为可以用来建构一个综合分析体系的概念轴心。这种‘宏大概念’突然流行起来，一时间把几乎所有的东西都挤到了一边。

这段话通常被认为是对当时“存在主义”和“精神分析法”这类万能概念的善意批评，而如今特别适合作为一盆冷水泼在那些没有任何深刻理解，却月月日日分分秒秒穿行于各种“大数据嘉年华”的投资人、媒体人和创业者身上。

希望《大数据时代》给予各位的是一些实实在在的知识和思考，并且唤起各位安静思索相关问题的心境。大数据是一个很重要的概念，代表了很重要的趋势，但我不希望它成为一种放之四海皆准的万能概念——因为越是万能的，就越是空洞的！人类学家克利福德·吉尔兹（Clifford Geertz）在其著作《文化的解释》中曾给出了一个朴素而冷静的劝说：“努力在可以应用、可以拓展的地方，应用它、拓展它；在不能应用、不能拓展的地方，就停下来。”我想，这应该是所有人面对一个新领域或新概念时应有的态度。

大数据的道路上没有戈多，我们已经在路上，晃晃悠悠。人类的自由意志和诸神之下的尊严，会在这条道路上异化甚至消逝吗？极目远眺，不知道世界的尽头，是否是一个冷酷的仙境！诸位为之奋斗吧，而我只想，做一个，麦田里的守望者。

以为序。

引言 一场生活、工作与思维的大变革

大数据开启了一次重大的时代转型。就像望远镜让我们能够感受宇宙，显微镜让我们能够观测微生物一样，大数据正在改变我们的生活以及理解世界的方式，成为新发明和新服务的源泉，而更多的改变正蓄势待发……

【大数据先锋】

谷歌搜索与流感预测

Farecast与飞机票价预测系统

天文学，信息爆炸的起源

大数据，变革公共卫生

2009年出现了一种新的流感病毒。这种甲型H1N1流感结合了导致禽流感和猪流感的病毒的特点，在短短几周之内迅速传播开来。全球的公共卫生机构都担心一场致命的流行病即将来袭。有的评论家甚至警告说，可能会爆发大规模流感，类似于1918年在西班牙爆发的影响了5亿人口并夺走了数千万人性命的大规模流感。更糟糕的是，我们还没有研发出对抗这种新型流感病毒的疫苗。公共卫生专家能做的只是减慢它传播的速度。但要做到这一点，他们必须先知道这种流感出现在哪里。

美国，和所有其他国家一样，都要求医生在发现新型流感病例时告知疾病控制与预防中心。但由于人们可能患病多日实在受不了了才会去医院，同时这个信息传达回疾控中心也需要时间，因此，通告新流感病例时往往会有一两周的延迟。而且，疾控中心每周只进行一次数据汇总。然而，对于一种飞速传播的疾病，信息滞后两周的后果将是致命的。这种滞后导致公共卫生机构在疫情爆发的关键时期反而无所适从。

在甲型H1N1流感爆发的几周前，互联网巨头谷歌公司的工程师们在《自然》杂志上发表了一篇引人注目的论文。它令公共卫生官员们和计算机科学家们感到震惊。文中解释了谷歌为什么能够预测冬季流感的传播：不仅是全美范围的传播，而且可以具体到特定的地区和州。谷歌通过观察人们在网上的搜索记录来完成这个预测，而这种方法以前一直是被忽略的。谷歌保存了多年来所有的搜索记录，而且每天都会收到来自全球超过30亿条的搜索指令，如此庞大的数据资源足以支撑和帮助它完成这项工作。

谷歌公司把5000万条美国人最频繁检索的词条和美国疾控中心在2003年至2008年间季节性流感传播时期的数据进行了比较。他们希望通过分析人们的搜索记录来判断这些人是否患上了流感，其他公司也曾试图确定这些相关的词条，但是他们缺乏像谷歌公司一样庞大的数据资源、处理能力和统计技术。

虽然谷歌公司的员工猜测，特定的检索词条是为了在网络上得到关于流感的信息，如“哪些是治疗咳嗽和发热的药物”，但是找出这些词条并不是重点，他们也不知道哪些词条更重要。更关键的是，他们建立的系统并不依赖于这样的语义理解。他们设立的这个系统唯一关注的就是

特定检索词条的使用频率与流感在时间和空间上的传播之间的联系。谷歌公司为了测试这些检索词条，总共处理了4.5亿个不同的数学模型。在将得出的预测与2007年、2008年美国疾控中心记录的实际流感病例进行对比后，谷歌公司发现，他们的软件发现了45条检索词条的组合，将它们用于一个特定的数学模型后，他们的预测与官方数据的相关性高达97%。和疾控中心一样，他们也能判断出流感是从哪里传播出来的，而且判断非常及时，不会像疾控中心一样要在流感爆发一两周之后才可以做到。

所以，2009年甲型H1N1流感爆发的时候，与习惯性滞后的官方数据相比，谷歌成为了一个更有效、更及时的指示标。公共卫生机构的官员获得了非常有价值的数据信息。惊人的是，谷歌公司的方法甚至不需要分发口腔试纸和联系医生——它是建立在大数据的基础之上的。这是当今社会所独有的一种新型能力：以一种前所未有的方式，通过对海量数据进行分析，获得有巨大价值的产品和服务，或深刻的洞见。基于这样的技术理念和数据储备，下一次流感来袭的时候，世界将会拥有一种更好的预测工具，以预防流感的传播。

大数据，变革商业

大数据不仅改变了公共卫生领域，整个商业领域都因为大数据而重新洗牌。购买飞机票就是一个很好的例子。

2003年，奥伦·埃齐奥尼（Oren Etzioni）准备乘坐从西雅图到洛杉矶的飞机去参加弟弟的婚礼。他知道飞机票越早预订越便宜，于是他在这个大喜日子来临之前的几个月，就在网上预订了一张去洛杉矶的机票。在飞机上，埃齐奥尼好奇地问邻座的乘客花了多少钱购买机票。当得知虽然那个人的机票比他买得更晚，但是票价却比他便宜得多时，他感到非常气愤。于是，他又询问了另外几个乘客，结果发现大家买的票居然都比他的便宜。

对大多数人来说，这种被敲竹杠的感觉也许会随着他们走下飞机而消失。然而，埃齐奥尼是美国最有名的计算机专家之一，从他担任华盛顿大学人工智能项目的负责人开始，他创立了许多在今天看来非常典型的大数据公司，而那时候还没有人提出“大数据”这个概念。

1994年，埃齐奥尼帮助创建了最早的互联网搜索引擎MetaCrawler，该引擎后来被InfoSpace公司收购。他联合创立了第一个大型比价网站Netbot，后来把它卖给了Excite公司。他创立的从文本中挖掘信息的公司ClearForest则被路透社收购了。在他眼中，世界就是一系列的大数据问题，而且他认为自己有能力解决这些问题。作为哈佛大学首届计算机科学专业的本科毕业生，自1986年毕业以来，他也一直致力于解决这些问题。

飞机着陆之后，埃齐奥尼下定决心要帮助人们开发一个系统，用来推测当前网页上的机票价格是否合理。作为一种商品，同一架飞机上每个座位的价格本来不应该有差别。但实际上，价格却千差万别，其中缘由只有航空公司自己清楚。

埃齐奥尼表示，他不需要去解开机票价格差异的奥秘。他要做的仅仅是预测当前的机票价格在未来一段时间内会上涨还是下降。这个想法是可行的，但操作起来并不是那么简单。这个系统需要分析所有特定航线机票的销售价格并确定票价与提前购买天数的关系。

如果一张机票的平均价格呈下降趋势，系统就会帮助用户做出稍后

再购票的明智选择。反过来，如果一张机票的平均价格呈上涨趋势，系统就会提醒用户立刻购买该机票。换言之，这是埃齐奥尼针对9000米高空开发的一个加强版的信息预测系统。这确实是一个浩大的计算机科学项目。不过，这个项目是可行的。于是，埃齐奥尼开始着手启动这个项目。

埃齐奥尼创立了一个预测系统，它帮助虚拟的乘客节省了很多钱。这个预测系统建立在41天之内的12000个价格样本基础之上，而这些数据都是从一个旅游网站上爬取过来的。这个预测系统并不能说明原因，只能推测会发生什么。也就是说，它不知道是哪些因素导致了机票价格的波动。机票降价是因为有很多没卖掉的座位、季节性原因，还是所谓的“周六晚上不出门”，它都不知道。这个系统只知道利用其他航班的数据来预测未来机票价格的走势。“买还是不买，这是一个问题。”埃齐奥尼沉思着。他给这个研究项目取了一个非常贴切的名字，叫“哈姆雷特”。

这个小项目逐渐发展成为一家得到了风险投资基金支持的科技创业公司，名为Farecast。通过预测机票价格的走势以及增降幅度，Farecast票价预测工具能帮助消费者抓住最佳购买时机，而在此之前还没有其他网站能让消费者获得这些信息。

这个系统为了保障自身的透明度，会把对机票价格走势预测的可信度标示出来，供消费者参考。系统的运转需要海量数据的支持。为了提高预测的准确性，埃齐奥尼找到了一个行业机票预订数据库。而系统的预测结果是根据美国商业航空产业中，每一条航线上每一架飞机内的每一个座位一年内的综合票价记录而得出的。如今，Farecast已经拥有惊人的约2000亿条飞行数据记录。利用这种方法，Farecast为消费者节省了一大笔钱。[1]

棕色的头发，露齿的笑容，无邪的面孔，这就是奥伦·埃齐奥尼。他看上去完全不像是一个会让航空业损失数百万潜在收入的人。但事实上，他的目光放得更长远。2008年，埃齐奥尼计划将这项技术应用到其他领域，比如宾馆预订、二手车购买等。只要这些领域内的产品差异不大，同时存在大幅度的价格差和大量可运用的数据，就都可以应用这项技术。但是在他实现计划之前，微软公司找上了他并以1.1亿美元的价格收购了Farecast公司。而后，这个系统被并入必应搜索引擎。

大数据的力量

到2012年为止，Farecast系统用了将近十万亿条价格记录来帮助预测美国国内航班的票价。Farecast票价预测的准确度已经高达75%，使用Farecast票价预测工具购买机票的旅客，平均每张机票可节省50美元。

Farecast是大数据公司的一个缩影，也代表了当今世界发展的趋势。五年或者十年之前，奥伦·埃齐奥尼是无法成立这样的公司的。他说："这是不可能的。"那时候他所需要的计算机处理能力和存储能力太昂贵了！虽说技术上的突破是这一切得以发生的主要原因，但也有一些细微而重要的改变正在发生，特别是人们关于如何使用数据的理念。

---

[1] 有趣的是，这些飞行记录和谷歌的搜索记录一样，也可以用来预测和评估疾病的流行。有兴趣的读者可以参考2010年第12期《科学通报》上名为"H1N1甲型流感全球航空传播与早期预警研究"的研究论文以及2011年Bajardi等人在*PLoS ONE*上发表的名为"Human Mobility Networks，Travel Restrictions，and the Global Spread of 2009 H1N1 Pandemic"的研究论文。——译者注

大数据，变革思维

人们不再认为数据是静止和陈旧的。但在以前，一旦完成了收集数据的目的之后，数据就会被认为已经没有用处了。比方说，在飞机降落之后，票价数据就没有用了（对谷歌而言，则是一个检索命令完成之后）。[1]

大数据洞察

如今，数据已经成为了一种商业资本，一项重要的经济投入，可以创造新的经济利益。事实上，一旦思维转变过来，数据就能被巧妙地用来激发新产品和新型服务。数据的奥妙只为谦逊、愿意聆听且掌握了聆听手段的人所知。

信息社会所带来的好处是显而易见的：每个人口袋里都揣有一部手机，每台办公桌上都放有一台电脑，每间办公室内都拥有一个大型局域网。但是，信息本身的用处却并没有如此引人注目。半个世纪以来，随着计算机技术全面融入社会生活，信息爆炸已经积累到了一个开始引发变革的程度。它不仅使世界充斥着比以往更多的信息，而且其增长速度也在加快。信息总量的变化还导致了信息形态的变化——量变引发了质变。最先经历信息爆炸的学科，如天文学和基因学，创造出了“大数据”这个概念[2]。如今，这个概念几乎应用到了所有人类致力于发展的领域中。

大数据并非一个确切的概念。最初，这个概念是指需要处理的信息量过大，已经超出了一般电脑在处理数据时所能使用的内存量，因此工程师们必须改进处理数据的工具。这导致了新的处理技术的诞生，例如谷歌的MapReduce和开源Hadoop平台（最初源于雅虎）。这些技术使得人们可以处理的数据量大大增加。更重要的是，这些数据不再需要用传统的数据库表格来整齐地排列——一些可以消除僵化的层次结构和一致性[3]的技术也出现了。同时，因为互联网公司可以收集大量有价值的数据，而且有利用这些数据的强烈的利益驱动力，所以互联网公司顺理成章地成为了最新处理技术的领头实践者。它们甚至超过了很多有几十年经验的线下公司，成为新技术的领衔使用者。

今天，一种可能的方式是，亦是本书采取的方式，认为大数据是人们在大规模数据的基础上可以做到的事情，而这些事情在小规模数据的

基础上是无法完成的。大数据是人们获得新的认知、创造新的价值的源泉；大数据还是改变市场、组织机构，以及政府与公民关系的方法。

大数据洞察

这仅仅只是一个开始，大数据时代对我们的生活，以及与世界交流的方式都提出了挑战。最惊人的是，社会需要放弃它对因果关系的渴求，而仅需关注相关关系。也就是说只需要知道是什么，而不需要知道为什么。这就推翻了自古以来的惯例，而我们做决定和理解现实的最基本方式也将受到挑战。

---

[1]　设计人员如果没有大数据的理念，就会丢失掉很多有价值的数据。譬如中国某城市的公交车因为价格不依赖于起点和终点，所以能够反映重要通勤信息的数据被工作人员“自作主张”地丢弃了。——译者注

[2]　有兴趣的读者可以参考2008年9月4日《自然》推出的名为“大数据”的专刊。——译者注

[3]　这些都是传统数据库结构化查询语言（SQL）的要求，非关系型数据库（NoSQL）不再有这些要求。——译者注

大数据，开启重大的时代转型

大数据开启了一次重大的时代转型。与其他新技术一样，大数据也必然要经历硅谷臭名昭著的技术成熟度曲线[1]：经过新闻媒体和学术会议的大肆宣传之后，新技术趋势一下子跌到谷底，许多数据创业公司变得岌岌可危。当然，不管是过热期还是幻想破灭期，都非常不利于我们正确理解正在发生的变革的重要性。

就像望远镜能够让我们感受宇宙，显微镜能够让我们观测微生物，这种能够收集和分析海量数据的新技术将帮助我们更好地理解世界——这种理解世界的新方法我们现在才意识到。本书旨在如实表达出大数据的内涵，而不会过分热捧它。当然，真正的革命并不在于分析数据的机器，而在于数据本身和我们如何运用数据。

大数据先锋

天文学，信息爆炸的起源

只有考虑到社会各个方面的变化趋势，我们才能真正意识到信息爆炸已经到来。我们的数字世界一直在扩张。以天文学为例，2000年斯隆数字巡天（Sloan Digital Sky Survey）项目启动的时候，位于新墨西哥州的望远镜在短短几周内收集到的数据，已经比天文学历史上总共收集的数据还要多。到了2010年，信息档案已经高达1.4×2^{42}字节。不过，预计2016年在智利投入使用的大型视场全景巡天望远镜（Large Synoptic Survey Telescope）能在五天之内就获得同样多的信息。

天文学领域的变化在各个领域都在发生。2003年，人类第一次破译人体基因密码的时候，辛苦工作了十年才完成了三十亿对碱基对的排序。大约十年之后，世界范围内的基因仪每15分钟就可以完成同样的工作。在金融领域，美国股市每天的成交量高达70亿股，而其中三分之二的交易都是由建立在数学模型和算法之上的计算机程序自动完成的。这些程序运用海量数据来预测利益和降低风险。

互联网公司更是要被数据淹没了。谷歌公司每天要处理超过24拍字节[2]的数据，这意味着其每天的数据处理量是美国国家图书馆所有纸质出版物所含数据量的上千倍。Facebook这个创立时间不足十年的公司，每天更新的照片量超过1000万张，每天人们在网站上点击“喜

欢”（Like）按钮或者写评论大约有三十亿次，这就为Facebook公司挖掘用户喜好提供了大量的数据线索。与此同时，谷歌子公司YouTube每月接待多达8亿的访客，平均每一秒钟就会有一段长度在一小时以上的视频上传。Twitter上的信息量几乎每年翻一番，截止到2012年，每天都会发布超过4亿条微博。

从科学研究到医疗保险，从银行业到互联网，各个不同的领域都在讲述着一个类似的故事，那就是爆发式增长的数据量。这种增长超过了我们创造机器的速度，甚至超过了我们的想象。

我们周围到底有多少数据？增长的速度有多快？许多人试图测量出一个确切的数字。尽管测量的对象和方法有所不同，但他们都获得了不同程度的成功。南加利福尼亚大学安嫩伯格通信学院的马丁·希尔伯特（Martin Hilbert）进行了一个比较全面的研究，他试图得出人类所创造、存储和传播的一切信息的确切数目。他的研究范围不仅包括书籍、图画、电子邮件、照片、音乐、视频（模拟和数字），还包括电子游戏、电话、汽车导航和信件。马丁·希尔伯特还以收视率和收听率为基础，对电视、电台这些广播媒体进行了研究。

大数据的力量

据他估算，2007年，人类大约存储了超过300艾字节[3]的数据。下面这个比喻应该可以帮助人们更容易地理解这意味着什么了。一部完整的数字电影可以压缩成一个GB的文件，而一个艾字节相当于10亿GB，一个泽字节[4]则相当于1024艾字节。总之，这是一个非常庞大的数量。

有趣的是，在2007年，所有数据中只有7%是存储在报纸、书籍、图片等媒介上的模拟数据[5]，其余全部是数字数据[6]。但在不久之前，情况却完全不是这样的。虽然1960年就有了“信息时代”和“数字村镇”的概念，但实际上，这些概念仍然是相当新颖的。甚至在2000年的时候，数字存储信息仍只占全球数据量的四分之一；当时，另外四分之三的信息都存储在报纸、胶片、黑胶唱片和盒式磁带这类媒介上。

早期数字信息的数量是不多的。对于长期在网上冲浪和购书的人来说，那只是一个微小的部分。事实上，在1986年的时候，世界上约40%的计算能力都被运用在袖珍计算器上，那时候，所有个人电脑的处理能力之和还没有所有袖珍计算器处理能力之和高。但是因为数字数据的快速增长，整个局势很快就颠倒过来了。按照希尔伯特的说法，数字数据

的数量每三年多就会翻一倍。相反，模拟数据的数量则基本上没有增加。

大数据的力量

到2013年，世界上存储的数据预计能达到约1.2泽字节，其中非数字数据只占不到2%。

这样大的数据量意味着什么？如果把这些数据全部记在书中，这些书可以覆盖整个美国52次。如果将之存储在只读光盘上，这些光盘可以堆成五堆，每一堆都可以伸到月球。公元前3世纪，埃及的托勒密二世竭力收集了当时所有的书写作品，所以伟大的亚历山大图书馆可以代表世界上所有的知识量。[7]但当数字数据洪流席卷世界之后，每个地球人都可以获得大量数据信息，相当于当时亚历山大图书馆存储的数据总量的320倍之多。

事情真的在快速发展。人类存储信息量的增长速度比世界经济的增长速度快**4**倍，而计算机数据处理能力的增长速度则比世界经济的增长速度快**9**倍。难怪人们会抱怨信息过量，因为每个人都受到了这种极速发展的冲击。

把眼光放远一点，我们可以把时下的信息洪流与1439年前后古登堡发明印刷机[8]时造成的信息爆炸相对比。历史学家伊丽莎白·爱森斯坦（Elizabeth Eisenstein）发现，1453—1503年，这50年之间大约有800万本书籍被印刷，比1200年之前君士坦丁堡建立以来整个欧洲所有的手抄书还要多。换言之，欧洲的信息存储量花了50年才增长了一倍（当时的欧洲还占据了世界上相当部分的信息存储份额），而如今大约每三年就能增长一倍。

这种增长意味着什么呢？彼特·诺维格（Peter Norvig）是谷歌的人工智能专家，也曾任职于美国宇航局喷气推进实验室，他喜欢把这种增长与图画进行类比。首先，他要我们想想来自法国拉斯科洞穴壁画上的标志性的马。这些画可以追溯到一万七千年之前的旧石器时代。然后，想想一张马的照片，再想想毕加索的画也可以，看起来和那些洞穴壁画没有多大的差别。事实上，毕加索看到那些洞穴壁画的时候就曾开玩笑说："自那以后，我们就再也没有创造出什么东西了。"

他的话既正确又不完全正确。你回想一下壁画上的那匹马。当时要

画一幅马需要花费很久的时间，而现在不需要那么久了。这就是一种改变，虽然改变的可能不是最核心的部分——毕竟这仍然是一幅马的图像。但是诺维格说，想象一下，现在我们能每秒钟播放24幅不同形态的马的图片，这就是一种由量变导致的质变：一部电影与一幅静态的画有本质上的区别！大数据也一样，量变导致质变。物理学和生物学都告诉我们，当我们改变规模时，事物的状态有时也会发生改变。

我们就以纳米技术为例。纳米技术专注于把东西变小而不是变大。其原理就是当事物到达分子的级别时，它的物理性质就会发生改变。一旦你知道这些新的性质，你就可以用同样的原料来做以前无法做的事情。铜本来是用来导电的物质，但它一旦到达纳米级别就不能在磁场中导电了。银离子具有抗菌性，但当它以分子形式存在的时候，这种性质会消失。一旦到达纳米级别，金属可以变得柔软，陶土可以具有弹性。同样，当我们增加所利用的数据量时，我们就可以做很多在小数据量的基础上无法完成的事情。

有时候，我们认为约束我们生活的那些限制，对于世间万物都有着同样的约束力。事实上，尽管规律相同，但是我们能够感受到的约束，很可能只对我们这样尺度的事物起作用。对于人类来说，唯一一个最重要的物理定律便是万有引力定律。这个定律无时无刻不在控制着我们。但对于细小的昆虫来说，重力是无关紧要的。[9]对它们而言，物理宇宙中有效的约束是表面张力，这个张力可以让它们在水上自由行走而不会掉下去。但人类对于表面张力毫不在意。

对于万有引力产生的约束效果而言，生物体的大小是非常重要的。类似地，对于信息而言，规模也是非常重要的。谷歌能够几近完美地给出和基于大量真实病例信息所得到的流感情况一致的结果，而且几乎是实时的，比疾控中心快多了。同样，Farecast可以预测机票价格的波动，从而让消费者真正在经济上获利。它们之所以如此给力，都因为存在供其分析的数千亿计的数据项。

大数据洞察

大数据的科学价值和社会价值正是体现在这里。一方面，对大数据的掌握程度可以转化为经济价值的来源。另一方面，大数据已经撼动了世界的方方面面，从商业科技到医疗、政府、教育、经济、人文以及社会的其他各个领域。

尽管我们仍处于大数据时代来临的前夕，但我们的日常生活已经离不开它了。垃圾邮件过滤器可以自动过滤垃圾邮件，尽管它并不知道“发#票#销#售”是“发票销售”的一种变体。交友网站根据个人的性格与之前成功配对的情侣之间的关联来进行新的配对。具有“自动改正”功能的智能手机通过分析我们以前的输入，将个性化的新单词添加到手机词典里。然而，对于这些数据的利用还仅仅只是一个开始。从可以自动转弯和刹车的汽车，到IBM沃森超级电脑在游戏节目《危险边缘》（*Jeopardy*）中打败人类来看，这项技术终将改变我们所居住的星球上的许多东西。

---

[1] 技术成熟度曲线又叫技术循环曲线，或者直接叫做炒作周期，是指新技术、新概念在媒体上曝光度随时间的变化曲线。——译者注

[2] 拍字节，一般记作PB，等于2^{50}字节。——译者注

[3] 艾字节，一般记作EB，等于2^{60}字节。——译者注

[4] 泽字节，一般记作ZB，等于2^{70}字节。——译者注

[5] 模拟数据也称为模拟量，相对于数字量而言，指的是取值范围是连续的变量或者数值，例如声音、图像、温度、压力等。模拟数据一般采用模拟信号，例如用一系列连续变化的电磁波或电压信号来表示。——译者注

[6] 数字数据也称为数字量，相对于模拟量而言，指的是取值范围是离散的变量或者数值。数字数据则采用数字信号，例如用一系列断续变化的电压脉冲（如用恒定的正电压表示二进制数1，用恒定的负电压表示二进制数0）或光脉冲来表示。——译者注

[7] 亚历山大图书馆藏书丰富，有据可考的超过50000卷（纸草卷），包括《荷马史诗》、《几何原本》等。亚历山大图书馆建成之时正是中国战国时代的末期，此时百家争鸣，较有影响的十大家（儒、道、墨、法、名、阴阳、纵横、杂、农、小说）多有著述，且已出现如《诗经》、《楚辞》、《离骚》等文学作品，虽没有像亚历山大图书馆一样的集中式藏书中心，但也占据了世界知识量的相当份额。——译者注

[8] 据《中国出版史》记载，中国的毕昇早在11世纪40年代就发明

了泥活字印刷，远远早于古登堡15世纪30年代发明的铅活字。——编者注

[9]　这是一个美妙有趣的例子，但是对于学习物理的人来说总是有些怪异。显然，万有引力一如既往起着作用，不过是因为空气阻力在不同密度和体积的物体上产生了不同的效果。如果把蟑螂从真空环境的高楼往下扔，恐怕也是凶多吉少。——译者注

预测，大数据的核心

大数据的核心就是预测。它通常被视为人工智能的一部分，或者更确切地说，被视为一种机器学习。但是这种定义是有误导性的。大数据不是要教机器像人一样思考。相反，它是把数学算法运用到海量的数据上来预测事情发生的可能性。一封邮件被作为垃圾邮件过滤掉的可能性，输入的“teh”应该是“the”的可能性，从一个人乱穿马路时行进的轨迹和速度来看他能及时穿过马路的可能性，都是大数据可以预测的范围。当然，如果一个人能及时穿过马路，那么他乱穿马路时，车子就只需要稍稍减速就好。这些预测系统之所以能够成功，关键在于它们是建立在海量数据的基础之上的。此外，随着系统接收到的数据越来越多，它们可以聪明到自动搜索最好的信号和模式，并自己改善自己。[1]

在不久的将来，世界许多现在单纯依靠人类判断力的领域都会被计算机系统所改变甚至取代。计算机系统可以发挥作用的领域远远不止驾驶和交友，还有更多更复杂的任务。别忘了，亚马逊可以帮我们推荐想要的书，谷歌可以为关联网站排序，Facebook知道我们的喜好，而LinkedIn可以猜出我们认识谁。[2]当然，同样的技术也可以运用到疾病诊断、推荐治疗措施，甚至是识别潜在犯罪分子上。

就像互联网通过给计算机添加通信功能而改变了世界，大数据也将改变我们生活中最重要的方面，因为它为我们的生活创造了前所未有的可量化的维度。大数据已经成为了新发明和新服务的源泉，而更多的改变正蓄势待发。

---

[1]　系统可以通过一种“反馈学习”的机制，利用自己产生的数据判断自身算法和参数选择的有效性，并实时进行调整，持续改进自身的表现。——译者注

[2]　这些任务都和个性化技术相关，包括个性化排序和个性化推荐。个性化技术是大数据时代最重要的技术，这里向专业读者推荐吕琳媛等人2012年在《Physics Reports》上发表的名为“Recommender Systems”的综述。——译者注

大数据，大挑战

大数据的精髓在于我们分析信息时的三个转变，这些转变将改变我们理解和组建社会的方法。

第一个转变就是，在大数据时代，我们可以分析更多的数据，有时候甚至可以处理和某个特别现象相关的所有数据，而不再依赖于随机采样。这部分内容将在第1章阐述。19世纪以来，当面临大量数据时，社会都依赖于采样分析。但是采样分析是信息缺乏时代和信息流通受限制的模拟数据时代的产物。以前我们通常把这看成是理所当然的限制，但高性能数字技术的流行让我们意识到，这其实是一种人为的限制。与局限在小数据范围相比，使用一切数据为我们带来了更高的精确性，也让我们看到了一些以前无法发现的细节——大数据让我们更清楚地看到了样本无法揭示的细节信息。

第二个改变就是，研究数据如此之多，以至于我们不再热衷于追求精确度。这部分内容将在第2章阐述。当我们测量事物的能力受限时，关注最重要的事情和获取最精确的结果是可取的。如果购买者不知道牛群里有80头牛还是100头牛，那么交易就无法进行。直到今天，我们的数字技术依然建立在精准的基础上。我们假设只要电子数据表格把数据排序，数据库引擎就可以找出和我们检索的内容完全一致的检索记录。

这种思维方式适用于掌握“小数据量”的情况，因为需要分析的数据很少，所以我们必须尽可能精准地量化我们的记录。在某些方面，我们已经意识到了差别。例如，一个小商店在晚上打烊的时候要把收银台里的每分钱都数清楚，但是我们不会、也不可能用“分”这个单位去精确度量国民生产总值。随着规模的扩大，对精确度的痴迷将减弱。

达到精确需要有专业的数据库。针对小数据量和特定事情，追求精确性依然是可行的，比如一个人的银行账户上是否有足够的钱开具支票。但是，在这个大数据时代，很多时候，追求精确度已经变得不可行，甚至不受欢迎了。当我们拥有海量即时数据时，绝对的精准不再是我们追求的主要目标。

大数据纷繁多样，优劣掺杂，分布在全球多个服务器上。拥有了大数据，我们不再需要对一个现象刨根究底，只要掌握大体的发展方向即

可。当然，我们也不是完全放弃了精确度，只是不再沉迷于此。适当忽略微观层面上的精确度会让我们在宏观层面拥有更好的洞察力。

第三个转变因前两个转变而促成，即我们不再热衷于寻找因果关系。这部分内容将在第3章阐述。寻找因果关系是人类长久以来的习惯。即使确定因果关系很困难而且用途不大，人类还是习惯性地寻找缘由。相反，在大数据时代，我们无须再紧盯事物之间的因果关系，而应该寻找事物之间的相关关系，这会给我们提供非常新颖且有价值的观点。相关关系也许不能准确地告知我们某件事情为何会发生，但是它会提醒我们这件事情正在发生。在许多情况下，这种提醒的帮助已经足够大了。

如果数百万条电子医疗记录显示橙汁和阿司匹林的特定组合可以治疗癌症，那么找出具体的药理机制就没有这种治疗方法本身来得重要。同样，只要我们知道什么时候是买机票的最佳时机，就算不知道机票价格疯狂变动的原因也无所谓了。大数据告诉我们“是什么”而不是“为什么”。在大数据时代，我们不必知道现象背后的原因，我们只要让数据自己发声。

我们不再需要在还没有收集数据之前，就把我们的分析建立在早已设立的少量假设的基础之上。让数据发声，我们会注意到很多以前从来没有意识到的联系的存在。

例如，对冲基金通过剖析社交网络Twitter上的数据信息来预测股市的表现；亚马逊和奈飞（Netflix）[1]根据用户在其网站上的类似查询来进行产品推荐；Twitter，Facebook和LinkedIn通过用户的社交网络图来得知用户的喜好。

当然，人类从数千年前就开始分析数据。古代美索不达米亚平原的记账人员为了有效地跟踪记录信息发明了书写。自从圣经时代开始，政府就通过进行人口普查来建立大型的国民数据库。两百多年来，精算师们也一直通过搜集大量的数据来进行风险规避。

模拟时代的数据收集和分析极其耗时耗力，新问题的出现通常要求我们重新收集和分析数据。数字化的到来使得数据管理效率又向前迈出了重要的一步。数字化将模拟数据转换成计算机可以读取的数字数据，使得存储和处理这些数据变得既便宜又容易，从而大大提高了数据管理效率。过去需要几年时间才能完成的数据搜集，现在只要几天就能完

成。但是，光有改变还远远不够。数据分析者太沉浸于模拟数据时代的设想，即数据库只有单一的用途和价值，而正是我们使用的技术和方法加深了这种偏见。虽然数字化是促成向大数据转变的重要原因，但仅有计算机的存在却不足以实现大数据。

我们没有办法准确描述现在正在发生的一切，但是在第4章将提到的“数据化”概念可以帮助我们大致了解这次变革。数据化意味着我们要从一切太阳底下的事物中汲取信息，甚至包括很多我们以前认为和“信息”根本搭不上边的事情。比方说，一个人所在的位置、引擎的振动、桥梁的承重等。我们要通过量化的方法把这些内容转化为数据。这就使得我们可以尝试许多以前无法做到的事情，如根据引擎的散热和振动来预测引擎是否会出现故障。这样，我们就激发出了这些数据此前未被挖掘的潜在价值。

大数据时代开启了一场寻宝游戏，而人们对于数据的看法以及对于由因果关系向相关关系转化时释放出的潜在价值的态度，正是主宰这场游戏的关键。新兴技术工具的使用使这一切成为可能。宝贝不止一件，每个数据集内部都隐藏着某些未被发掘的价值。这场发掘和利用数据价值的竞赛正开始在全球上演。

第5章和第6章将讲述大数据如何改变了商业、市场和社会的本质。20世纪，价值已经从实体基建转变为无形财产，从土地和工厂转变为品牌和产权。如今，一个新的转变正在进行，那就是电脑存储和分析数据的方法取代电脑硬件成为了价值的源泉。数据成为了有价值的公司资产、重要的经济投入和新型商业模式的基石。虽然数据还没有被列入企业的资产负债表，但这只是一个时间问题。

虽然有些数据处理技术已经出现了一段时间，但是它们只为调查局、研究所和世界上的一些巨头公司所掌握。沃尔玛和美国第一资本银行（CapitalOne）率先将大数据运用在了零售业和银行业，因此改变了整个行业。如今这些技术大多都实现了大众化。

大数据对个人的影响是最惊人的。在一个可能性和相关性占主导地位的世界里，专业性变得不那么重要了。行业专家不会消失，但是他们必须与数据表达的信息进行博弈。如同在电影《点球成金》（*Moneyball*）里，棒球星探们在统计学家面前相形见绌——直觉的判断被迫让位于精准的数据分析。这将迫使人们调整在管理、决策、人力资源和教育方面的传统理念。

我们大部分的习俗和惯例都建立在一个预设好的立场上，那就是我们用来进行决策的信息必须是少量、精确并且至关重要的。但是，当数据量变大、数据处理速度加快，而且数据变得不那么精确时，之前的那些预设立场就不复存在了。此外，因为数据量极为庞大，最后做出决策的将是机器而不是人类自己。第7章将会讨论大数据的负面影响。

在了解和监视人类的行为方面，社会已经有了数千年的经验。但是，如何来监管一个算法系统呢？在信息化时代的早期，有一些政策专家就看到了信息化给人们的隐私权带来的威胁，社会也已经建立起了庞大的规则体系来保障个人的信息安全。但是在大数据时代，这些规则都成了无用的马其诺防线[2]。人们自愿在网络上分享信息，而这种分享的能力成为了网络服务的一个中心特征，而不再是一个需要规避的薄弱点了。

对我们而言，危险不再是隐私的泄露，而是被预知的可能性——这些能预测我们可能生病、拖欠还款和犯罪的算法会让我们无法购买保险、无法贷款、甚至在实施犯罪前就被预先逮捕。显然，统计把大数据放在了首位，但即便如此，个人意志是否应该凌驾于大数据之上呢？就像出版印刷行业的发展推动国家立法保护言论自由（在此之前没有出台类似法律的必要，因为没有太多的言论需要保护），大数据时代也需要新的规章制度来保卫权势面前的个人权利。

政府机构和社会在控制和处理数据的方法上必须有全方位的改变。不可否认，我们进入了一个用数据进行预测的时代，虽然我们可能无法解释其背后的原因。如果一个医生只要求病人遵从医嘱，却没法说明医学干预的合理性的话，情况会怎么样呢？实际上，这是依靠大数据取得病理分析的医生们一定会做的事情。还有司法系统的“合理证据”是不是应该改为“可能证据”呢？如果真是这样，会对人类自由和尊严产生什么影响呢？

我们在大数据时代倡导的一系列规范将在第8章进行介绍。这些规范建立在我们很熟悉的“小数据”时代发展并保留下来的规范的基础之上。新环境要求旧规范与时俱进。

大数据洞察

大数据给社会带来的益处将是多方面的。因为大数据已经成为解决紧迫世界性问题，如抑制全球变暖、消除疾病、提高执政能力和发展经

济的一个有力武器。但是大数据时代也向我们提出了挑战，我们需要做好充足的准备迎接大数据技术给我们的机构和自身带来的改变。

大数据标志着人类在寻求量化和认识世界的道路上前进了一大步。过去不可计量、存储、分析和共享的很多东西都被数据化了。拥有大量的数据和更多不那么精确的数据为我们理解世界打开了一扇新的大门。社会因此放弃了寻找因果关系的传统偏好，开始挖掘相关关系的好处。

寻找原因是一种现代社会的一神论，大数据推翻了这个论断。但我们又陷入了一个历史的困境，那就是我们活在一个“上帝已死”的时代。也就是说，我们曾经坚守的信念动摇了。讽刺的是，这些信念正在被“更好”的证据所取代。那么，从经验中得来的与证据相矛盾的直觉、信念和迷惘应该充当什么角色呢？当世界由探求因果关系变成挖掘相关关系，我们怎样才能既不损坏建立在因果推理基础之上的社会繁荣和人类前行的基石，又取得实际的进步呢？本书意在解释我们身在何处，我们从何而来，并且提供当下亟需的指导，以应对眼前的利益和危险。

---

[1] Netflix，也常译作网狸公司。——译者注

[2] 马奇诺防线是法国在第一次世界大战后，为防德军入侵而在其东北边境地区构筑的筑垒配系，以其陆军部长姓氏命名。1940年5月至6月，德国主力通过阿登山脉，从马奇诺防线左翼迂回，进抵马奇诺防线的后方，使防线丧失了作用。“马奇诺防线”现在用来意指看似表面坚固，实际毫无价值的东西。——译者注

第一部分 大数据时代的思维变革

01 更多：不是随机样本，而是全体数据

当数据处理技术已经发生了翻天覆地的变化时，在大数据时代进行抽样分析就像在汽车时代骑马一样。一切都改变了，我们需要的是所有的数据，“样本=总体”。

【大数据先锋】

穿孔卡片与美国人口普查

大数据与乔布斯的癌症治疗

Xoom与跨境汇款异常交易报警

巴拉巴西与第一次全社会层面的网络分析

让数据“发声”

“大数据”全在于发现和理解信息内容及信息与信息之间的关系，然而直到最近，我们对此似乎还是难以把握。IBM的资深“大数据”专家杰夫·乔纳斯（Jeff Jonas）提出要让数据“说话”。从某种层面上来说，这听起来很平常。人们使用数据已经有相当长一段时间了，无论是日常进行的大量非正式观察，还是过去几个世纪里在专业层面上用高级算法进行的量化研究，都与数据有关。

在数字化时代，数据处理变得更加容易、更加快速，人们能够在瞬间处理成千上万的数据。但当我们谈论能“说话”的数据时，我们指的远远不止这些。

实际上，大数据与三个重大的思维转变有关，这三个转变是相互联系和相互作用的。

●首先，要分析与某事物相关的所有数据，而不是依靠分析少量的数据样本。

●其次，我们乐于接受数据的纷繁复杂，而不再追求精确性。

●最后，我们的思想发生了转变，不再探求难以捉摸的因果关系，转而关注事物的相关关系。

本章就将介绍第一个转变：利用所有的数据，而不再仅仅依靠一小部分数据。

很长一段时间以来，准确分析大量数据对我们而言都是一种挑战。过去，因为记录、储存和分析数据的工具不够好，我们只能收集少量数据进行分析，这让我们一度很苦恼。为了让分析变得简单，我们会把数据量缩减到最少。这是一种无意识的自省：我们把与数据交流的困难看成是自然的，而没有意识到这只是当时技术条件下的一种人为的限制。如今，技术条件已经有了非常大的提高，虽然人类可以处理的数据依然是有限的，也永远是有限的，但是我们可以处理的数据量已经大大地增加，而且未来会越来越多。

在某些方面，我们依然没有完全意识到自己拥有了能够收集和处理更大规模数据的能力。我们还是在信息匮乏的假设下做很多事情，建立很多机构组织。我们假定自己只能收集到少量信息，结果就真的如此了。这是一个自我实现的过程。我们甚至发展了一些使用尽可能少的信息的技术。别忘了，统计学的一个目的就是用尽可能少的数据来证实尽可能重大的发现。事实上，我们形成了一种习惯，那就是在我们的制度、处理过程和激励机制中尽可能地减少数据的使用。为了理解大数据时代的转变意味着什么，我们需要首先回顾一下过去。

小数据时代的随机采样，最少的数据获得最多的信息

直到最近，私人企业和个人才拥有了大规模收集和分类数据的能力。在过去，这是只有教会或者政府才能做到的。当然，在很多国家，教会和政府是等同的。有记载的、最早的计数发生在公元前8000年的，当时苏美尔的商人用黏土珠来记录出售的商品。大规模的计数则是政府的事情。数千年来，政府都试图通过收集信息来管理国民。

以人口普查为例。据说古代埃及曾进行过人口普查，《旧约》和《新约》中对此都有所提及。那次由奥古斯都恺撒主导实施的人口普查，提出了“每个人都必须纳税”，这使得约瑟夫和玛丽搬到了耶稣的出生地伯利恒。1086年的《末日审判书》（*The Doomsday Book*）对当时

英国的人口、土地和财产做了一个前所未有的全面记载。皇家委员穿越整个国家对每个人、每件事都做了记载，后来这本书用《圣经》中的《末日审判书》命名，因为每个人的生活都被赤裸裸地记载下来的过程就像接受“最后的审判”一样。

然而，人口普查是一项耗资且费时的事情。国王威廉一世（King William I）在他发起的《末日审判书》完成之前就去世了。但是，除非放弃收集信息，否则在当时没有其他办法。尽管如此，当时收集的信息也只是一个大概情况，实施人口普查的人也知道他们不可能准确记录下每个人的信息。实际上，“人口普查”这个词来源于拉丁语的“*censere*”，意思就是推测、估算。

三百多年前，一个名叫约翰·格朗特（John Graunt）[1]的英国缝纫用品商提出了一个很有新意的方法。他采用了一个新方法推算出鼠疫时期伦敦的人口数，这种方法就是后来的统计学。这个方法不需要一个人一个人地计算。虽然这个方法比较粗糙，但采用这个方法，人们可以利用少量有用的样本信息来获取人口的整体情况。

虽然后来证实他能够得出正确的数据仅仅是因为运气好，但在当时他的方法大受欢迎。样本分析法一直都有较大的漏洞，因此无论是进行人口普查还是其他大数据类的任务，人们还是一直使用一一清点这种“野蛮”的方法。

考虑到人口普查的复杂性以及耗时耗费的特点，政府极少进行普查。古罗马在拥有数十万人口的时候每5年普查一次。美国宪法规定每10年进行一次人口普查，而随着国家人口越来越多，只能以百万计数。但是到19世纪为止，即使这样不频繁的人口普查依然很困难，因为数据变化的速度超过了人口普查局统计分析的能力。

大数据先锋

穿孔卡片与美国人口普查

美国在1880年进行的人口普查，耗时8年才完成数据汇总。因此，他们获得的很多数据都是过时的。1890年进行的人口普查，预计要花费13年的时间来汇总数据。即使不考虑这种情况违反了宪法规定，它也是很荒谬的。然而，因为税收分摊和国会代表人数确定都是建立在人口的基础上的，所以必须要得到正确的数据，而且必须是及时的数据。

美国人口普查局面临的问题与当代商人和科学家遇到的问题很相似。很明显，当他们被数据淹没的时候，已有的数据处理工具已经难以应付了，所以就需要有更多的新技术。

后来，美国人口普查局就和当时的美国发明家赫尔曼·霍尔瑞斯（Herman Hollerith）签订了一个协议，用他的穿孔卡片制表机来完成1890年的人口普查。

经过大量的努力，霍尔瑞斯成功地在1年时间内完成了人口普查的数据汇总工作。这简直就是一个奇迹，它标志着自动处理数据的开端，也为后来IBM公司的成立奠定了基础。但是，将其作为收集处理大数据的方法依然过于昂贵。毕竟，每个美国人都必须填一张可制成穿孔卡片的表格，然后再进行统计。这么麻烦的情况下，很难想象如果不足十年就要进行一次人口普查应该怎么办。但是，对于一个跨越式发展的国家而言，十年一次的人口普查的滞后性已经让普查失去了大部分意义。

这就是问题所在，是利用所有的数据还是仅仅采用一部分呢？最明智的自然是得到有关被分析事物的所有数据，但是当数量无比庞大时，这又不太现实。那如何选择样本呢？有人提出有目的地选择最具代表性的样本是最恰当的方法。1934年，波兰统计学家耶日·奈曼（Jerzy Neyman）指出，这只会导致更多更大的漏洞。事实证明，问题的关键是选择样本时的随机性。[2]

统计学家们证明：采样分析的精确性随着采样随机性的增加而大幅提高，但与样本数量的增加关系不大。虽然听起来很不可思议，但事实上，一个对1100人进行的关于“是否”问题的抽样调查有着很高的精确性，精确度甚至超过了对所有人进行调查时的97%[3]。这是真的，不管是调查10万人还是1亿人，20次调查里有19次都能猜对。为什么会这样？原因很复杂，但是有一个比较简单的解释就是，当样本数量达到了某个值之后，我们从新个体身上得到的信息会越来越少，就如同经济学中的边际效应递减一样。

认为样本选择的随机性比样本数量更重要，这种观点是非常有见地的。这种观点为我们开辟了一条收集信息的新道路。通过收集随机样本，我们可以用较少的花费做出高精准度的推断。因此，政府每年都可以用随机采样的方法进行小规模的人口普查，而不是只能每十年进行一次。事实上，政府也这样做了。例如，除了十年一次的人口大普查，美国人口普查局每年都会用随机采样的方法对经济和人口进行200多次小

规模的调查。当收集和分析数据都不容易时，随机采样就成为应对信息采集困难的办法。

很快，随机采样就不仅应用于公共部门和人口普查了。在商业领域，随机采样被用来监管商品质量。这使得监管商品质量和提升商品品质变得更容易，花费也更少。以前，全面的质量监管要求对生产出来的每个产品进行检查，而现在只需从一批商品中随机抽取部分样品进行检查就可以了。本质上来说，随机采样让大数据问题变得更加切实可行。同理，它将客户调查引进了零售行业，将焦点讨论引进了政治界，也将许多人文问题变成了社会科学问题。

随机采样取得了巨大的成功，成为现代社会、现代测量领域的主心骨。但这只是一条捷径，是在不可收集和分析全部数据的情况下的选择，它本身存在许多固有的缺陷。[4]它的成功依赖于采样的绝对随机性，但是实现采样的随机性非常困难。一旦采样过程中存在任何偏见，分析结果就会相去甚远。

最近，以固定电话用户为基础进行投票民调就面临了这样的问题，采样缺乏随机性，因为没有考虑到只使用移动电话的用户——这些用户一般更年轻和更热爱自由。没有考虑到这些用户，自然就得不到正确的预测。2008年在奥巴马与麦凯恩之间进行的美国总统大选中，盖洛普咨询公司、皮尤研究中心（Pew）、美国广播公司和《华盛顿邮报》社这些主要的民调组织都发现，如果他们不把移动用户考虑进来，民意测试结果就会出现三个点的偏差，而一旦考虑进来，偏差就只有一个点。鉴于这次大选的票数差距极其微弱，这已经是非常大的偏差了。

更糟糕的是，随机采样不适合考察子类别的情况。因为一旦继续细分，随机采样结果的错误率会大大增加。这很容易理解。倘若你有一份随机采样的调查结果，是关于1000个人在下一次竞选中的投票意向。如果采样时足够随机，这份调查的结果就有可能在3%的误差范围内显示全民的意向。但是如果这个3%左右的误差本来就是不确定的，却又把这个调查结果根据性别、地域和收入进行细分，结果是不是越来越不准确呢？用这些细分过后的结果来表现全民的意愿，是否合适呢？

你设想一下，一个对1000个人进行的调查，如果要细分到“东北部的富裕女性”，调查的人数就远远少于1000人了。即使是完全随机的调查，倘若只用了几十个人来预测整个东北部富裕女性选民的意愿，还是不可能得到精确结果啊！而且，一旦采样过程中存在任何偏见，在细分

领域所做的预测就会大错特错。

因此，当人们想了解更深层次的细分领域的情况时，随机采样的方法就不可取了。在宏观领域起作用的方法在微观领域失去了作用。随机采样就像是模拟照片打印，远看很不错，但是一旦聚焦某个点，就会变得模糊不清。

随机采样也需要严密的安排和执行。人们只能从采样数据中得出事先设计好的问题的结果——千万不要奢求采样的数据还能回答你突然意识到的问题。所以虽说随机采样是一条捷径，但它也只是一条捷径。随机采样方法并不适用于一切情况，因为这种调查结果缺乏延展性，即调查得出的数据不可以重新分析以实现计划之外的目的。

我们来看一下DNA分析。由于技术成本大幅下跌以及在医学方面的广阔前景，个人基因排序成为了一门新兴产业。2012年，基因组解码的价格跌破1000美元，这也是非正式的行业平均水平。从2007年起，硅谷的新兴科技公司23andme就开始分析人类基因，价格仅为几百美元。这可以揭示出人类遗传密码中一些会导致其对某些疾病抵抗力差的特征，如乳腺癌和心脏病。23andme希望能通过整合顾客的DNA和健康信息，了解到用其他方式不能获取的新信息。

公司对某人的一小部分DNA进行排序，标注出几十个特定的基因缺陷。这只是该人整个基因密码的样本，还有几十亿个基因碱基对未排序。最后，23andme只能回答其标注过的基因组表现出来的问题。发现新标注时，该人的DNA必须重新排序，更准确地说，是相关的部分必须重新排列。只研究样本而不是整体，有利有弊：能更快更容易地发现问题，但不能回答事先未考虑到的问题。

大数据先锋

大数据与乔布斯的癌症治疗

苹果公司的传奇总裁史蒂夫·乔布斯在与癌症斗争的过程中采用了不同的方式，成为世界上第一个对自身所有DNA和肿瘤DNA进行排序的人。为此，他支付了高达几十万美元的费用，这是23andme报价的几百倍之多。所以，他得到的不是一个只有一系列标记的样本，他得到了包括整个基因密码的数据文档。

对于一个普通的癌症患者，医生只能期望她的DNA排列同试验中使用的样本足够相似。但是，史蒂夫·乔布斯的医生们能够基于乔布斯的特定基因组成，按所需效果用药。如果癌症病变导致药物失效，医生可以及时更换另一种药，也就是乔布斯所说的，“从一片睡莲叶跳到另一片上。”乔布斯开玩笑说：“我要么是第一个通过这种方式战胜癌症的人，要么就是最后一个因为这种方式死于癌症的人。”虽然他的愿望都没有实现，但是这种获得所有数据而不仅是样本的方法还是将他的生命延长了好几年。

全数据模式，样本=总体

在信息处理能力受限的时代，世界需要数据分析，却缺少用来分析所收集数据的工具，因此随机采样应运而生，它也可以被视为那个时代的产物。如今，计算和制表不再像过去一样困难。感应器、手机导航、网站点击和Twitter被动地收集了大量数据，而计算机可以轻易地对这些数据进行处理。

采样的目的就是用最少的数据得到最多的信息。当我们可以获得海量数据的时候，它就没有什么意义了。数据处理技术已经发生了翻天覆地的改变，但我们的方法和思维却没有跟上这种改变。

采样一直有一个被我们广泛承认却又总有意避开的缺陷，现在这个缺陷越来越难以忽视了。采样忽视了细节考察。虽然我们别无选择，只能利用采样分析法来进行考察，但是在很多领域，从收集部分数据到收集尽可能多的数据的转变已经发生了。如果可能的话，我们会收集所有的数据，即“样本=总体”。

正如我们所看到的，“样本=总体”是指我们能对数据进行深度探讨，而采样几乎无法达到这样的效果。上面提到的有关采样的例子证明，用采样的方法分析整个人口的情况，正确率可达97%。对于某些事物来说，3%的错误率是可以接受的。但是你无法得到一些微观细节的信息，甚至还会失去对某些特定子类别进行进一步研究的能力。我们不能满足于正态分布一般中庸平凡的景象。生活中真正有趣的事情经常藏匿在细节之中，而采样分析法却无法捕捉到这些细节。

谷歌流感趋势预测并不是依赖于对随机样本的分析，而是分析了整个美国几十亿条互联网检索记录。分析整个数据库，而不是对一个小样本进行分析，能够提高微观层面分析的准确性，甚至能够推测出某个特

定城市的流感状况，而不只是一个州或是整个国家的情况。Farecast的初始系统使用的样本包含12000个数据，所以取得了不错的预测结果。随着奥伦·埃齐奥尼不断添加更多的数据，预测的结果越来越准确。最终，Farecast使用了每一条航线整整一年的价格数据来进行预测。埃齐奥尼说："这只是一个暂时性的数据，随着你收集的数据越来越多，你的预测结果会越来越准确。"

所以，我们现在经常会放弃样本分析这条捷径，选择收集全面而完整的数据。我们需要足够的数据处理和存储能力，也需要最先进的分析技术。同时，简单廉价的数据收集方法也很重要。过去，这些问题中的任何一个都很棘手。在一个资源有限的时代，要解决这些问题需要付出很高的代价。但是现在，解决这些难题已经变得简单容易得多。曾经只有大公司才能做到的事情，现在绝大部分的公司都可以做到了。

通过使用所有的数据，我们可以发现如若不然则将会在大量数据中淹没掉的情况。例如，信用卡诈骗是通过观察异常情况来识别的，只有掌握了所有的数据才能做到这一点。在这种情况下，异常值是最有用的信息，你可以把它与正常交易情况进行对比。这是一个大数据问题。而且，因为交易是即时的，所以你的数据分析也应该是即时的。

大数据先锋

Xoom与跨境汇款异常交易报警

Xoom是一个专门从事跨境汇款业务的公司，它得到了很多拥有大数据的大公司的支持。它会分析一笔交易的所有相关数据。2011年，它注意到用"发现卡"从新泽西州汇款的交易量比正常情况多一些，系统于是启动报警。Xoom公司的首席执行官约翰·孔泽（John Kunze）解释说："这个系统关注的是不应该出现的情况。"单独来看，每笔交易都是合法的，但是事实证明这是一个犯罪集团在试图诈骗。而发现异常的唯一方法就是，重新检查所有的数据，找出样本分析法错过的信息。

然而，使用所有的数据并不代表这是一项艰巨的任务。大数据中的"大"不是绝对意义上的大，虽然在大多数情况下是这个意思。谷歌流感趋势预测建立在数亿的数学模型上，而它们又建立在数十亿数据节点的基础之上。完整的人体基因组有约30亿个碱基对。但这只是单纯的数据节点的绝对数量，并不代表它们就是大数据。大数据是指不用随机分析法这样的捷径，而采用所有数据的方法。谷歌流感趋势和乔布斯的

医生们采取的就是大数据的方法。

日本国民体育运动“相扑”中非法操纵比赛结果的发现过程，就恰到好处地说明了使用“样本=总体”这种全数据模式的重要性。消极比赛一直被极力禁止，备受谴责，很多运动员深受困扰。芝加哥大学的一位很有前途的经济学家斯蒂夫·列维特（Steven Levitt），在《美国经济评论》上发表了一篇研究论文，其中提到了一种发现这种情况的方法：查看运动员过去所有的比赛资料。他的畅销书《魔鬼经济学》（*Freakonomics*）中也提到了这个观点，他认为检查所有的数据是非常有价值的。

列维特和他的同事马克·达根（Mark Duggan）使用了11年中超过64000场摔跤比赛的记录，来寻找异常性。他们获得了重大的发现。非法操纵比赛结果的情况确实时有发生，但是不会出现在大家很关注的比赛上。冠军赛也有可能被操纵，但是数据显示消极比赛主要还是出现在不太被关注的联赛的后几场中。这时基本上没有什么风险，因为很多选手根本就没有获奖的希望。

相扑比赛的一个比较特殊的地方是，选手需要在15场赛事中的大部分场次取得胜利才能保持排名和收入。这样一来就会出现利益不对称的问题。当一名7胜7负的摔跤手碰到一个8胜6负的对手时，比赛结果对第一个选手来说极其重要，对他的对手而言则没有那么重要。列维特和达根发现，在这样的情况下，需要赢的那个选手很可能会赢。这看起来像是对手送的“礼物”，因为在联系紧密的相扑界，帮别人一把就是给自己留了一条后路。

有没有可能是要赢的决心帮助这个选手获胜呢？答案是，有可能。但是数据显示的情况是，需要赢的选手的求胜心也只能把胜率提高25%。所以，把胜利完全归功于求胜心是不妥当的。对数据进行进一步分析可能会发现，与他们在先前比赛中的表现相比，当他们再相遇时，上次失利的一方要拥有比对方更高的胜率。因为在相扑界，你的付出总会有所“回报”，所以第一次的胜利看上去更像是一名选手送给另一名选手的礼物。

这个情况是显而易见的。但是如果采用随机采样分析法，就无法发现这个情况。而大数据分析通过使用所有比赛的极大数据捕捉到了这个情况。这就像捕鱼一样，开始时你不知道是否能捕到鱼，也不知道会捕到什么鱼。

一个数据库并不需要有以太字节[5]计的数据。在这个相扑案例中，整个数据库包含的字节量还不如一张普通的数码照片包含得多。但是大数据分析法不只关注一个随机的样本。这里的"大"取的是相对意义而不是绝对意义，也就是说这是相对所有数据来说的。

很长一段时间内，随机采样都是一条好的捷径，它使得数字时代之前的大量数据分析变得可能。但就像把一张数码照片或者一首数码歌曲截取成多个小文件似的，在采样分析的时候，很多信息就丢失了——你能欣赏一首歌的抽样吗？拥有全部或几乎全部的数据，我们就能够从不同的角度，更细致地观察和研究数据的方方面面。

我们可以用Lytro相机来打一个恰当的比方。Lytro相机是具有革新性的，因为它把大数据运用到了基本的摄影中。与传统相机只可以记录一束光不同，Lytro相机可以记录整个光场里所有的光，达到1100万束之多。具体生成什么样的照片则可以在拍摄之后再根据需要决定。用户没必要在一开始就聚焦，因为该相机可以捕捉到所有的数据，所以之后可以选择聚焦图像中的任一点。整个光场的光束都被记录了，也就是收集了所有的数据，"样本=总体"。因此，与普通照片相比，这些照片就更具"可循环利用性"。如果使用普通相机，摄影师就必须在拍照之前决定好聚焦点。

同理，因为大数据是建立在掌握所有数据，至少是尽可能多的数据的基础上的，所以我们就可以正确地考察细节并进行新的分析。在任何细微的层面，我们都可以用大数据去论证新的假设。是大数据让我们发现了相扑中的非法操纵比赛结果、流感的传播区域和对抗癌症需要针对的那部分DNA。它让我们能清楚分析微观层面的情况。

当然，有些时候，我们还是可以使用样本分析法，毕竟我们仍然活在一个资源有限的时代。但是更多时候，利用手中掌握的所有数据成为了最好也是可行的选择。

社会科学是被"样本=总体"撼动得最厉害的学科。随着大数据分析取代了样本分析，社会科学不再单纯依赖于分析实证数据。这门学科过去曾非常依赖样本分析、研究和调查问卷。当记录下来的是人们的平常状态，也就不用担心在做研究和调查问卷时存在的偏见[6]了。现在，我们可以收集过去无法收集到的信息，不管是通过移动电话表现出的关系，还是通过Twitter信息表现出的感情。更重要的是，我们现在也不再依赖抽样调查了。

艾伯特-拉斯洛·巴拉巴西（Albert-László Barabási）[7]，和他的同事想研究人与人之间的互动。于是他们调查了四个月内所有的移动通信记录——当然是匿名的，这些记录是一个为全美五分之一人口提供服务的无线运营商提供的。这是第一次在全社会层面用接近于“样本=总体”的数据资料进行网络分析。通过观察数百万人的所有通信记录，我们可以产生也许通过任何其他方式都无法产生的新观点。

有趣的是，与小规模的研究相比，这个团队发现，如果把一个在社区内有很多连接关系的人从社区关系网中剔除掉，这个关系网会变得没那么高效但却不会解体；但如果把一个与所在社区之外的很多人有着连接关系的人从这个关系网中剔除，整个关系网很快就会破碎成很多小块。[8]这个研究结果非常重要也非常得出人意料。谁能想象一个在关系网内有着众多好友的人的重要性还不如一个只是与很多关系网外的人有联系的人呢？[9]这说明一般来说无论是针对一个小团体还是整个社会，多样性是有额外价值的。这个结果促使我们重新审视一个人在社会关系网中的存在价值。

大数据洞察

我们总是习惯把统计抽样看做文明得以建立的牢固基石，就如同几何学定理和万有引力定律一样。但是统计抽样其实只是为了在技术受限的特定时期，解决当时存在的一些特定问题而产生的，其历史尚不足一百年。如今，技术环境已经有了很大的改善。在大数据时代进行抽样分析就像是在汽车时代骑马一样。在某些特定的情况下，我们依然可以使用样本分析法，但这不再是我们分析数据的主要方式。慢慢地，我们会完全抛弃样本分析。

---

[1] 约翰·格朗特的尝试可以参见他闻名世界的著作*Natural and Political Observations Made upon the Bills of Mortality*。尽管他并未真正给出一种有效的办法来推断疾病流行时的人口数或死亡率，但是他首次建立了区分各年龄段的存活率表，因此被认为是人口统计学的主要创始人之一。——译者注

[2] 在对一个量（例如年收入）进行估计的时候，如果总体可以分为很多层（例如所有人口按照不同年龄或者不同职业分成很多层），一种直观的想法是每一层随机抽样的样本大小应该正比于这个层所包含人口的多少。奈曼指出，最优分配并非如此简单，实际上，层越大，层内

待估计量的变化越大，该层抽样的单位费用越小，则该层的抽样就应该越多。具体的公式和推导过程可以参考1934年奈曼在*Journal of the Royal Statistical Society*上发表的“On the two different aspects of the representative method：The method of stratified sampling and the method of purposive selection”一文。——译者注。

[3] 此处指对回答“是”的概率的估计精度可以达到97%左右，也就是说估计值的标准差在3%左右。——译者注

[4] 刚才讨论的还只是最简单的经典抽样问题。尽管奈曼等人指出了其中非平凡的问题，但毕竟存在最优抽样的判断标准和最优方法。最近，祝建华教授在一次讨论中指出，如果抽样的对象更复杂，例如是一个网络，那么根本找不到一个“最优抽样”的标准，更不可能奢求抽样得到的小网络能够反映总体的所有结构特性。——译者注

[5] 太字节，一般记作TB，等于2^{40}字节。——译者注

[6] 这种偏见既包括研究者设计实验和问卷时的偏差，也包括被试人员由于了解自己作为被试的角色而产生的不同于日常的心理和行为。——译者注

[7] 全球最权威的复杂网络研究专家之一，其最新作品《爆发》（*Bursts*）是大数据新科学范式的典型代表，该书的简体中文版已由湛庐文化策划、中国人民大学出版社出版，推荐与本书参照阅读。——译者注

[8] 作者对这项研究的理解稍有不妥。该研究并未关注从网络中移除节点（手机用户）的情形，而是考察从网络中移除链路（通话关系）对网络结构的影响。借鉴渗流理论（Percolation Theory），作者发现，移除弱关系而非强关系反而会导致快速破碎成若干小碎片。详细分析可以参考Onnela等人2007年在《美国科学院院刊》上发表的“Structure and tie strengths in mobile communication networks”一文。——译者注

[9] 就个人而言，可以通过重叠社区挖掘的方法找到同时属于多个社区的节点，这些人往往对网络连通性至关重要。就联系而言，可以挖掘起桥接作用的连边，这些连边往往对网络连通性至关重要。这方面的概念和算术可参考2005年Palla等人在《自然》上发表的名为“Uncovering the overlapping community structure of complex networks in nature and

society”一文及2010年程学旗等人在《统计力学杂志》上发表的名为“Bridgeness：a local index on edge significance in maintaining global connectivity”一文。——译者注

02 更杂：不是精确性，而是混杂性

执迷于精确性是信息缺乏时代和模拟时代的产物。只有5%的数据是结构化且能适用于传统数据库的。如果不接受混乱，剩下95%的非结构化数据都无法被利用，只有接受不精确性，我们才能打开一扇从未涉足的世界的窗户。

【大数据先锋】

微软与语料库数据添加

IBM Candidate计算机翻译项目

无所不包的谷歌翻译系统

英国石油公司与无线感应器

麻省理工与通货紧缩预测软件

Hadoop与VISA的13分钟

允许不精确

在越来越多的情况下，使用所有可获取的数据变得更为可能，但为此也要付出一定的代价。数据量的大幅增加会造成结果的不准确，与此同时，一些错误的数据也会混进数据库。然而，重点是我们能够努力避免这些问题。我们从不认为这些问题是无法避免的，而且也正在学会接受它们。这就是由“小数据”到“大数据”的重要转变之一。

对“小数据”而言，最基本、最重要的要求就是减少错误，保证质量。因为收集的信息量比较少，所以我们必须确保记录下来的数据尽量精确。无论是确定天体的位置还是观测显微镜下物体的大小，为了使结果更加准确，很多科学家都致力于优化测量的工具。在采样的时候，对精确度的要求就更高更苛刻了。因为收集信息的有限意味着细微的错误会被放大，甚至有可能影响整个结果的准确性。

历史上很多时候，人们会把通过测量世界来征服世界视为最大的成

就。事实上，对精确度的高要求始于13世纪中期的欧洲。那时候，天文学家和学者对时间、空间的研究采取了比以往更为精确的量化方式，用历史学家阿尔弗雷德·克罗斯比（Alfred Crosby）的话来说就是“测量现实”。

我们研究一个现象，是因为我们相信我们能够理解它。后来，测量方法逐渐被运用到科学观察、解释方法中，体现为一种进行量化研究、记录，并呈现可重复结果的能力。伟大的物理学家开尔文男爵曾说过：“测量就是认知。”这已成为一条至理名言。培根也曾说过：“知识就是力量。”同时，很多数学家以及后来的精算师和会计师都发展了可以准确收集、记录和管理数据的方法。

19世纪，科技率先发展起来的法国开发了一套能准确计量时间、空间单位的系统，并逐渐成为其他国家普遍采用的标准，这套系统还为后来国际公认的测量条约奠定了基础，成为测量时代的巅峰。仅半个世纪之后，20世纪20年代，量子力学的发现永远粉碎了“测量臻于至善”的幻梦。然而，在物理学这个小圈子以外的一些测量工程师和科学家仍沉湎在完美测量的梦中。随着理性学科，如数学和统计学逐渐影响到商业领域，商业界更加崇尚这种思想。

然而，在不断涌现的新情况里，允许不精确的出现已经成为一个新的亮点，而非缺点。因为放松了容错的标准，人们掌握的数据也多了起来，还可以利用这些数据做更多新的事情。这样就不是大量数据优于少量数据那么简单了，而是大量数据创造了更好的结果。

同时，我们需要与各种各样的混乱做斗争。混乱，简单地说就是随着数据的增加，错误率也会相应增加。所以，如果桥梁的压力数据量增加1000倍的话，其中的部分读数就可能是错误的，而且随着读数量的增加，错误率可能也会继续增加。在整合来源不同的各类信息的时候，因为它们通常不完全一致，所以也会加大混乱程度。例如，与服务器处理投诉时的数据进行比较，用语音识别系统识别某个呼叫中心接到的投诉会产生一个不太准确的结果，但也是有助于我们把握整个事情的大致情况的。

混乱还可以指格式的不一致性，因为要达到格式一致，就需要在进行数据处理之前仔细地清洗数据，而这在大数据背景下很难做到。“大数据”专家帕堤尔（D.J.Patil）指出，I.B.M.、T.J.Watson Labs、International Business Machines都可以用来指代IBM，甚至可能有成千上

万种方法称呼IBM。当然，在萃取或处理数据的时候，混乱也会发生。因为在进行数据转化的时候，我们是在把它变成另外的事物。比如，我们在对Twitter的信息进行情感分析来预测好莱坞票房的时候，就会出现一定的混乱。其实，混乱的起源和类型本来就是一团乱麻。

假设你要测量一个葡萄园的温度，但是整个葡萄园只有一个温度测量仪，那你就必须确保这个测量仪是精确的而且能够一直工作。反过来，如果每100棵葡萄树就有一个测量仪，有些测试的数据可能会是错误的，可能会更加混乱，但众多的读数合起来就可以提供一个更加准确的结果。因为这里面包含了更多的数据，而它不仅能抵消掉错误数据造成的影响，还能提供更多的额外价值。

现在想想增加读数频率的这个事情。如果每隔一分钟就测量一下温度，我们至少还能够保证测量结果是按照时间有序排列的。如果变成每分钟测量十次甚至百次的话，不仅读数可能出错，连时间先后都可能搞混掉。试想，如果信息在网络中流动，那么一条记录很可能在传输过程中被延迟，在其到达的时候已经没有意义了，甚至干脆在奔涌的信息洪流中彻底迷失。虽然我们得到的信息不再那么准确，但收集到的数量庞大的信息让我们放弃严格精确的选择变得更为划算。

在第一个例子里，我们为了获得更广泛的数据而牺牲了精确性，也因此看到了很多如若不然无法被关注到的细节。在第二个例子里，我们为了高频率而放弃了精确性，结果观察到了一些本可能被错过的变化。虽然如果我们能够下足够多的工夫，这些错误是可以避免的，但在很多情况下，与致力于避免错误相比，对错误的包容会带给我们更多好处。

为了扩大规模，我们接受适量错误的存在。正如技术咨询公司Forrester所认为的，有时得到2加2约等于3.9的结果，也很不错了。当然，数据不可能完全错误，但为了了解大致的发展趋势，我们愿意对精确性做出一些让步。

大数据洞察

“大数据”通常用概率说话，而不是板着“确凿无疑”的面孔。整个社会要习惯这种思维需要很长的时间，其中也会出现一些问题。但现在，有必要指出的是，当我们试图扩大数据规模的时候，要学会拥抱混乱。

我们可以在大量数据对计算机其他领域进步的重要性上看到类似的

变化。我们都知道，如摩尔定律所预测的，过去一段时间里计算机的数据处理能力得到了很大的提高。摩尔定律认为，每块芯片上晶体管的数量每两年就会翻一倍。这使得电脑运行更快速了，存储空间更大了。大家没有意识到的是，驱动各类系统的算法也进步了——美国总统科技顾问委员会的报告显示，在很多领域这些算法带来的进步还要胜过芯片的进步。然而，社会从“大数据”中所能得到的，并非来自运行更快的芯片或更好的算法，而是更多的数据。

由于象棋的规则家喻户晓，且走子限制良多，在过去的几十年里，象棋算法的变化很小。计算机象棋程序总是步步为赢是由于对残局掌握得更好了，[1]而之所以能做到这一点也只是因为往系统里加入了更多的数据。实际上，当棋盘上只剩下六枚棋子或更少的时候，这个残局得到了全面的分析，并且接下来所有可能的走法（样本=总体）都被制入了一个庞大的数据表格[2]。这个数据表格如果不压缩的话，会有一太字节那么多。所以，计算机在这些重要的象棋残局中表现得完美无缺和不可战胜。

大数据在多大程度上优于算法这个问题在自然语言处理上表现得很明显（这是关于计算机如何学习和领悟我们在日常生活中使用语言的学科方向）。在2000年的时候，微软研究中心的米歇尔·班科（Michele Banko）和埃里克·布里尔（Eric Bill）一直在寻求改进Word程序中语法检查的方法。但是他们不能确定是努力改进现有的算法、研发新的方法，还是添加更加细腻精致的特点更有效。所以，在实施这些措施之前，他们决定往现有的算法中添加更多的数据，看看会有什么不同的变化。很多对计算机学习算法的研究都建立在百万字左右的语料库基础上。最后，他们决定往4种常见的算法中逐渐添加数据，先是一千万字，再到一亿字，最后到十亿。

结果有点令人吃惊。他们发现，随着数据的增多，4种算法的表现都大幅提高了。

大数据的力量

当数据只有500万的时候，有一种简单的算法表现得很差，但当数据达10亿的时候，它变成了表现最好的，准确率从原来的75%提高到了95%以上。与之相反地，在少量数据情况下运行得最好的算法，当加入更多的数据时，也会像其他的算法一样有所提高，但是却变成了在大量数据条件下运行得最不好的。它的准确率会从86%提高到94%。

后来，班科和布里尔在他们发表的研究论文中写到，“如此一来，我们得重新衡量一下更多的人力物力是应该消耗在算法发展上还是在语料库发展上。”

大数据的简单算法比小数据的复杂算法更有效

所以，数据多比少好，更多数据比算法系统更智能还要重要。那么，混乱呢？在班科和布里尔开始研究数据几年后，微软的最大竞争对手，谷歌，也开始更大规模地对这些问题进行探讨。谷歌用的是上万亿的语料库，而不是十亿的。谷歌做这类研究不是因为语法检查，而是为了解决翻译这个更棘手的难题。

20世纪40年代，电脑由真空管制成，要占据整个房间这么大的空间。而机器翻译也只是计算机开发人员的一个想法。在冷战时期，美国掌握了大量关于苏联的各种资料，但缺少翻译这些资料的人手。所以，计算机翻译也成了亟须解决的问题。

最初，计算机研发人员打算将语法规则和双语词典结合在一起。1954年，IBM以计算机中的250个词语和六条语法规则为基础，将60个俄语词组翻译成了英语，结果振奋人心。IBM 701通过穿孔卡片读取了“Mi pyeryedayem mislyi posryedstvom ryechyi”这句话，并且将其译成了“我们通过语言来交流思想”。在庆祝这个成就的发布会上，一篇报道就有提到，这60句话翻译得很流畅。这个程序的指挥官利昂·多斯特尔特（Leon Dostert）表示，他相信“在三五年后，机器翻译将会变得很成熟”。

事实证明，计算机翻译最初的成功误导了人们。1966年，一群机器翻译的研究人员意识到，翻译比他们想象的更困难，他们不得不承认自己的失败。机器翻译不能只是让电脑熟悉常用规则，还必须教会电脑处理特殊的语言情况。毕竟，翻译不仅仅只是记忆和复述，也涉及选词，而明确地教会电脑这些非常不现实。法语中的“bonjour”就一定是“早上好”吗？有没有可能是“今天天气不错”、“吃了吗”或者“喂”？事实上都有可能——这需要视情况而定。

在20世纪80年代后期，IBM的研发人员提出了一个新的想法。与单纯教给计算机语言规则和词汇相比，他们试图让计算机自己估算一个词或一个词组适合于用来翻译另一种语言中的一个词和词组的可能性，然后再决定某个词和词组在另一种语言中的对等词和词组。

20世纪90年代，IBM这个名为Candide的项目花费了大概十年的时间，将大约有300万句之多的加拿大议会资料译成了英语和法语并出版。由于是官方文件，翻译的标准就非常高。用那个时候的标准来看，数据量非常之庞大。统计机器学习从诞生之日起，就聪明地把翻译的挑战变成了一个数学问题，而这似乎很有效！计算机翻译能力在短时间内就提高了很多。然而，在这次飞跃之后，IBM公司尽管投入了很多资金，但取得的成效不大。最终，IBM公司停止了这个项目。

大数据先锋

无所不包的谷歌翻译系统

2006年，谷歌公司也开始涉足机器翻译。这被当作实现“收集全世界的数据资源，并让人人都可享受这些资源”这个目标的一个步骤。谷歌翻译开始利用一个更大更繁杂的数据库，也就是全球的互联网，而不再只利用两种语言之间的文本翻译。

谷歌翻译系统为了训练计算机，会吸收它能找到的所有翻译。它会从各种各样语言的公司网站上寻找对译文档，还会去寻找联合国和欧盟这些国际组织发布的官方文件和报告的译本。它甚至会吸收速读项目中的书籍翻译。谷歌翻译部的负责人弗朗兹·奥齐（Franz Och）是机器翻译界的权威，他指出，“谷歌的翻译系统不会像Candide一样只是仔细地翻译300万句话，它会掌握用不同语言翻译的质量参差不齐的数十亿页的文档。”不考虑翻译质量的话，上万亿的语料库就相当于950亿句英语。

尽管其输入源很混乱，但较其他翻译系统而言，谷歌的翻译质量相对而言还是最好的，而且可翻译的内容更多。到2012年年中，谷歌数据库涵盖了60多种语言，甚至能够接受14种语言的语音输入，并有很流利的对等翻译。之所以能做到这些，是因为它将语言视为能够判别可能性的数据，而不是语言本身。如果要将印度语译成加泰罗尼亚语，谷歌就会把英语作为中介语言。因为在翻译的时候它能适当增减词汇，所以谷歌的翻译比其他系统的翻译灵活很多。

谷歌的翻译之所以更好并不是因为它拥有一个更好的算法机制。和微软的班科和布里尔一样，这是因为谷歌翻译增加了很多各种各样的数据。从谷歌的例子来看，它之所以能比IBM的Candide系统多利用成千上万的数据，是因为它接受了有错误的数据。2006年，谷歌发布的上万

亿的语料库，就是来自于互联网的一些废弃内容。这就是“训练集”，可以正确地推算出英语词汇搭配在一起的可能性。

20世纪60年代，拥有百万英语单词的语料库——布朗语料库算得上这个领域的开创者，而如今谷歌的这个语料库则是一个质的突破，后者使用庞大的数据库使得自然语言处理这一方向取得了飞跃式的发展。自然语言处理能力是语音识别系统和计算机翻译的基础。

彼得·诺维格（Peter Norvig），谷歌公司人工智能方面的专家，和他的同事在一篇题为《数据的非理性效果》（*The Unreasonable Effectiveness of Data*）的文章中写道，“大数据基础上的简单算法比小数据基础上的复杂算法更加有效。”他们就指出，混杂是关键。

“从某种意义上，谷歌的语料库是布朗语料库的一个退步。因为谷歌语料库的内容来自于未经过滤的网页内容，所以会包含一些不完整的句子、拼写错误、语法错误以及其他各种错误。况且，它也没有详细的人工纠错后的注解。但是，谷歌语料库是布朗语料库的好几百万倍大，这样的优势完全压倒了缺点。”

纷繁的数据越多越好

传统的样本分析师们很难容忍错误数据的存在，因为他们一生都在研究如何防止和避免错误的出现。在收集样本的时候，统计学家会用一整套的策略来减少错误发生的概率。在结果公布之前，他们也会测试样本是否存在潜在的系统性偏差。这些策略包括根据协议或通过受过专门训练的专家来采集样本。但是，即使只是少量的数据，这些规避错误的策略实施起来还是耗费巨大。尤其是当我们收集所有数据的时候，这就行不通了。不仅是因为耗费巨大，还因为在大规模的基础上保持数据收集标准的一致性不太现实。就算是不让人们进行沟通争吵，也不能解决这个问题。

大数据时代要求我们重新审视精确性的优劣。如果将传统的思维模式运用于数字化、网络化的21世纪，就会错过重要的信息。执迷于精确性是信息缺乏时代和模拟时代的产物。在那个信息贫乏的时代，任意一个数据点的测量情况都对结果至关重要。所以，我们需要确保每个数据的精确性，才不会导致分析结果的偏差。

大数据洞察

如今，我们已经生活在信息时代。我们掌握的数据库越来越全面，它不再只包括我们手头现象的一点点可怜的数据，而是包括了与这些现象相关的大量甚至全部数据。我们不再需要那么担心某个数据点对整套分析的不利影响。我们要做的就是要接受这些纷繁的数据并从中受益，而不是以高昂的代价消除所有的不确定性。

在华盛顿州布莱恩市的英国石油公司（BP）切里波因特（Cherry Point）炼油厂里，无线感应器遍布于整个工厂，形成无形的网络，能够产生大量实时数据。酷热的恶劣环境和电气设备的存在有时会对感应器读数有所影响，形成错误的数据。但是数据生成的数量之多可以弥补这些小错误。随时监测管道的承压使得BP能够了解到，有些种类的原油比其他种类更具有腐蚀性。以前，这都是无法发现也无法防止的。

有时候，当我们掌握了大量新型数据时，精确性就不那么重要了，我们同样可以掌握事情的发展趋势。大数据不仅让我们不再期待精确性，也让我们无法实现精确性。然而，除了一开始会与我们的直觉相矛盾之外，接受数据的不精确和不完美，我们反而能够更好地进行预测，也能够更好地理解这个世界。

值得注意的是，错误性并不是大数据本身固有的。它只是我们用来测量、记录和交流数据的工具的一个缺陷。如果说哪天技术变得完美无缺了，不精确的问题也就不复存在了。错误并不是大数据固有的特性，而是一个亟需我们去处理的现实问题，并且有可能长期存在。因为拥有更大数据量所能带来的商业利益远远超过增加一点精确性，所以通常我们不会再花大力气去提升数据的精确性。这又是一个关注焦点的转变，正如以前，统计学家们总是把他们的兴趣放在提高样本的随机性而不是数量上。如今，大数据给我们带来的利益，让我们能够接受不精确的存在了。

大数据先锋

麻省理工与通货紧缩预测软件

“10亿价格项目”（The Billion Prices Project，BBP）提供了一个有趣的例子。美国劳工统计局的人员每个月都要公布消费物价指数（CPI），这是用来测试通货膨胀率的。这些数据对投资者和商家都非常重要。在决定是否增减银行利率的时候，美联储也会考虑消费指数。一旦发生通货膨胀，工人工资也会增加。联邦政府在支付社会福利和债

券利息的款项时，这项指数也是他们参考的依据。

联邦政府为了得到这些数据，会雇用很多人向全美90个城市的商店、办公室打电话、发传真甚至登门拜访。他们反馈回来的各种各样的价格信息达80000种，包括土豆的价格、出租车的票价等。政府采集这些数据每年大概需要花费两亿五千万美元。这些数据是精确的也是有序的，但是这个采集结果的公布会有几周的滞后。2008年的经济危机表明，这个滞后是致命的。政策决策者为了更好地应对变化，需要及时了解通货膨胀率，但如果以传统的依赖采样和追求精确的方式进行数据收集，政府就不可能及时获得数据了。

麻省理工学院（MIT）的两位经济学家，阿尔贝托·卡瓦略（Alberto Cavell）和罗伯托·里哥本（Oberto Rigobon）就对此提出了一个大数据方案，那就是接受更混乱的数据。通过一个软件在互联网上收集信息，他们每天可以收集到50万种商品的价格。收集到的数据很混乱，也不是所有数据都能轻易进行比较。但是把大数据和好的分析法相结合，这个项目在2008年9月雷曼兄弟破产之后马上就发现了通货紧缩趋势，然而那些依赖官方数据的人直到11月份才知道这个情况。[3]

MIT的这个项目汇集了数百万的产品，它们被数百个零售商卖到了70多个国家。这个项目产生的一个名为PriceStats的商业方案也经常被一些银行和其他经济决策人用到。当然，收集到的数据需要仔细的分析，而且这些数据更善于表明价格的发展趋势而不是精确的价格。但是因为PriceStats收集到了更多的价格信息而且大多是即时的，所以这对决策者来说就非常有益了。

混杂性，不是竭力避免，而是标准途径

确切地说，在许多技术和社会领域，我们更倾向于纷繁混杂。我们来看看内容分类方面的情况。几个世纪以来，人们一直用分类法和索引法来帮助自己存储和检索数据资源。这样的分级系统通常都不完善——各位读者没有忘记图书馆卡片目录给你们带来的痛苦回忆吧？在“小数据”范围内，这些方法就很有效，但一旦把数据规模增加好几个数量级，这些预设一切都各就各位的系统就会崩溃。

相片分享网站Flickr在2011年拥有来自大概1亿用户的60亿张照片。根据预先设定好的分类来标注每张照片就没有意义了。难道真会有人为他的照片取名“像希特勒一样的猫”吗？

恰恰相反，清楚的分类被更混乱却更灵活的机制所取代了。这些机制才能适应改变着的世界。当我们上传照片到Flickr网站的时候，我们会给照片添加标签。也就是说，我们会使用一组文本标签来编组和搜索这些资源。人们用自己的方式创造和使用标签，所以它是没有标准、没有预先设定的排列和分类，也没有我们必须遵守的类别的。任何人都可以输入新的标签，标签内容事实上就成为了网络资源的分类标准。标签被广泛地应用于Facebook、博客等社交网络上。因为它们的存在，互联网上的资源变得更加容易找到，特别是像图片、视频和音乐这些无法用关键词搜索的非文本类资源。[4]

当然，有时人们错标的标签会导致资源编组的不准确，这会让习惯了精确性的人们很痛苦。但是，我们用来编组照片集的混乱方法给我们带来了很多好处。比如，我们拥有了更加丰富的标签内容，同时能更深更广地获得各种照片。我们可以通过合并多个搜索标签来过滤我们需要寻找的照片，这在以前是无法完成的。我们添加标签时所固带的不准确性从某种意义上说明我们能够接受世界的纷繁复杂。这是对更加精确系统的一种对抗。这些精确的系统试图让我们接受一个世界贫乏而规整的惨象——假装世间万物都是整齐地排列的。而事实上现实是纷繁复杂的，天地间存在的事物也远远多于系统所设想的。

互联网上最火的网址都表明，它们欣赏不精确而不会假装精确。当一个人在网站上见到一个Facebook的“喜欢”按钮时，可以看到有多少其他人也在点击。当数量不多时，会显示像“63”这种精确的数字。当数量很大时，则只会显示近似值，比方说“4000”。这并不代表系统不知道正确的数据是多少，只是当数量规模变大的时候，确切的数量已经不那么重要了。另外，数据更新得非常快，甚至在刚刚显示出来的时候可能就已经过时了。所以，同样的原理适用于时间的显示。谷歌的Gmail邮箱会确切标注在很短时间内收到的信件，比方说“11分钟之前”。但是，对于已经收到一段时间的信件，则会标注如“两个小时之前”这种不太确切的时间信息。

2000年以来，商务智能和分析软件领域的技术供应商们一直承诺给客户“一个唯一的真理”。执行官们用这个词组并没有讽刺的意思，现在也依然有技术供应商这样说。他们说这个词组的意思就是，每个使用该公司信息技术系统的人都能利用同样的数据资源，这样市场部和营销部的人员们就不需要再在会议开始前争论，到底是谁掌握了正确的客户和销售数据了。这个想法就是说，如果他们知道的数据是一致的，那么他

们的利益也会更一致。

但是，“一个唯一的真理”这种想法已经彻底被改变了。现在不但出现了一种新的认识，即“一个唯一的真理”的存在是不可能的，而且追求这个唯一的真理是对注意力的分散。要想获得大规模数据带来的好处，混乱应该是一种标准途径，而不应该是竭力避免的。

我们甚至发现，不精确已经渗入了数据库设计这个最不能容忍错误的领域。传统的数据库引擎要求数据高度精确和准确排列。数据不是单纯地被存储，它往往被划分为包含“域”的记录，每个域都包含了特定种类和特定长度的信息。比方说，某个数值域是7个数字长，一个1000万或者更大的数值就无法被记录。一个人想在某个记录手机号码的域中输入一串汉字是“不被允许”的。想要被允许也可以，需要改变数据库结构才可以。现在，我们依然在和电脑以及智能手机上的这些限制进行斗争，比如软件可能拒绝记录我们输入的数据。

索引是事先就设定好了的，这也就限制了人们的搜索。增加一个新的索引往往既消耗时间，又惹人讨论，因为需要改变底层的设计。传统的关系数据库是为数据稀缺的时代设计的，所以能够也需要仔细策划。在那个时代，人们遇到的问题无比清晰，所以数据库被设计用来有效地回答这些问题。

但是，这种数据存储和分析的方法越来越和现实相冲突。我们现在拥有各种各样、参差不齐的海量数据。很少有数据完全符合预先设定的数据种类。而且，我们想要数据回答的问题，也只有在我们收集和处理数据的过程中才会知道。

新的数据库设计的诞生

这些现实条件导致了新的数据库设计的诞生，它们打破了关于记录和预设场域的成规。预设场域显示的是数据的整齐排列。最普遍的数据库查询语言是结构化查询语言，英文缩写为“SQL”——它的名字就显示了它的僵化。但是，近年的大转变就是非关系型数据库的出现，它不需要预先设定记录结构，允许处理超大量五花八门的数据。因为包容了结构多样性，这些数据库设计就要求更多的处理和存储资源。但是，一旦考虑到大大降低的存储和处理成本，这就是一个我们支付得起的公平交易。

帕特·赫兰德（Pat Helland）是来自微软的世界上最权威的数据库设计专家之一，在一篇题为《如果你有足够多的数据，那么“足够好”真的足够好》（*If You Have Too Much Data，then‘Good Enough’Is Good Enough*）的文章中，他把这称为一个重大的转变。分析了被各种各样质量参差不齐的数据所侵蚀的传统数据库设计的核心原则，他得出的结论是，“我们再也不能假装活在一个齐整的世界里”。他认为，处理海量数据会不可避免地导致部分信息的缺失。虽然这本来就是有“损耗性”的，但是能快速得到想要的结果弥补了这个缺陷。赫兰德总结说：“略有瑕疵的答案并不会伤了商家的胃口，因为他们更看重高频率。”

传统数据库的设计要求在不同的时间提供一致的结果。比方说，如果你查询你的账户结余，它会提供给你确切的数目；而你几秒钟之后查询的时候，系统应该提供给你同样的结果，没有任何改变。但是，随着数据数量的大幅增加以及系统用户的增加，这种一致性将越来越难保持。

大的数据库并不是固定在某个地方的，它一般分散在多个硬盘和多台电脑上。为了确保其运行的稳定性和速度，一个记录可能会分开存储在两三个地方。如果一个地方的记录更新了，其他地方的记录则只有同步更新才不会产生错误。传统的系统会一直等到所有地方的记录都更新，然而，当数据广泛地分布在多台服务器上而且服务器每秒钟都会接受成千上万条搜索指令的时候，同步更新就比较不现实了。因此，多样性是一种解决的方法。

大数据先锋

Hadoop与VISA的13分钟

最能代表这个转变的，就是Hadoop的流行。Hadoop是与谷歌的MapReduce系统相对应的开源式分布系统的基础架构，它非常善于处理超大量的数据。通过把大数据变成小模块然后分配给其他机器进行分析，它实现了对超大量数据的处理。它预设硬件可能会瘫痪，所以在内部建立了数据的副本，它还假定数据量之大导致数据在处理之前不可能整齐排列。典型的数据分析需要经过“萃取、转移和下载”这样一个操作流程，但是Hadoop就不拘泥于这样的方式。相反，它假定了数据量的巨大使得数据完全无法移动，所以人们必须在本地进行数据分析。

Hadoop的输出结果没有关系型数据库输出结果那么精确，它不能用

于卫星发射、开具银行账户明细这种精确度要求很高的任务。但是对于不要求极端精确的任务，它就比其他系统运行得快很多，比如说把顾客分群，然后分别进行不同的营销活动。

信用卡公司VISA使用Hadoop，能够将处理两年内730亿单交易所需的时间，从一个月缩减至仅仅13分钟。这样大规模处理时间上的缩减足以变革商业了。也许Hadoop不适合正规记账，但是当可以允许少量错误的时候它就非常实用。

ZestFinance，一个由谷歌前任首席信息官道格拉斯·梅里尔创立的公司，用自己的经验再次验证了“宽容错误会给我们带来更多价值”这一观点。这家公司帮助决策者判断是否应该向某些拥有不良信用记录的人提供小额短期贷款。传统的信用评分机制关注少量突出的事件，比如一次还款的延迟，而ZestFinance则分析了大量不那么突出的事件。2012年，让ZestFinance引以为豪的就是，它的贷款拖欠率比行业平均水平要低三分之一左右。唯一的得胜之道还是拥抱混杂。

梅里尔说：“有趣的是，对我们而言，基本没有任何一个人的信息是齐备的，事实上，总有大量的数据缺失。”由ZestFinance创建的用来记录客户信息的矩阵是难以想象得稀疏，里面充满了数据的空洞，但ZestFinance在这些支离破碎的数据中如鱼得水。举个例子，有10%的客户属性信息显示“已经死亡”，但是依然可以从他们身上收回贷款。梅里尔一脸坏笑地说：“显然，没有人会企盼僵尸复活并且主动还贷。但是我们的数据显示，放贷给僵尸是一项不错的生意。”

接受混乱，我们就能享受极其有用的服务，这些服务如果使用传统方法和工具是不可能做到的，因为那些方法和工具处理不了这么大规模的数据。

大数据的力量

据估计，只有5%的数字数据是结构化的且能适用于传统数据库。如果不接受混乱，剩下95%的非结构化数据都无法被利用，比如网页和视频资源。通过接受不精确性，我们打开了一个从未涉足的世界的窗户。

社会将两个折中的想法不知不觉地渗入了我们的处事方法中，我们甚至不再把这当成一种折中，而是把它当成了事物的自然状态。

第一个折中是，我们默认自己不能使用更多的数据，所以我们就不会去使用更多的数据。但是，数据量的限制正在逐渐消失，而且通过无限接近“样本=总体”的方式来处理数据，我们会获得极大的好处。

第二个折中出现在数据的质量上。在小数据时代，追求精确度是合理的。因为当时我们收集的数据很少，所以需要越精确越好。如今这依然适用于一些事情。但是对于其他事情，快速获得一个大概的轮廓和发展脉络，就要比严格的精确性要重要得多。

大数据洞察

我们怎么看待使用所有数据和使用部分数据的差别，以及我们怎样选择放松要求并取代严格的精确性，将会对我们与世界的沟通产生深刻的影响。随着大数据技术成为日常生活中的一部分，我们应该开始从一个比以前更大更全面的角度来理解事物，也就是说应该将“样本=总体”植入我们的思维中。

现在，我们能够容忍模糊和不确定出现在一些过去依赖于清晰和精确的领域，当然过去可能也只是有清晰的假象和不完全的精确。只要我们能够得到一个事物更完整的概念，我们就能接受模糊和不确定的存在。就像印象派的画风一样，近看画中的每一笔都感觉是混乱的，但是退后一步你就会发现这是一幅伟大的作品，因为你退后一步的时候就能看出画作的整体思路了。

相比依赖于小数据和精确性的时代，大数据因为更强调数据的完整性和混杂性，帮助我们进一步接近事实的真相。“部分”和“确切”的吸引力是可以理解的。但是，当我们的视野局限在我们可以分析和能够确定的数据上时，我们对世界的整体理解就可能产生偏差和错误。不仅失去了去尽力收集一切数据的动力，也失去了从各个不同角度来观察事物的权利。所以，局限于狭隘的小数据中，我们可以自豪于对精确性的追求，但是就算我们可以分析得到细节中的细节，也依然会错过事物的全貌。

大数据洞察

大数据要求我们有所改变，我们必须能够接受混乱和不确定性。精确性似乎一直是我们生活的支撑，就像我们常说的“钉是钉，铆是铆”。但认为每个问题只有一个答案的想法是站不住脚的，不管我们承不承

认。一旦我们承认了这个事实甚至拥护这个事实的话，我们离真相就又近了一步。

这些思想上的重大转变导致了第三个变革，这个变革有望颠覆很多传统观念。这些传统观念更加基本，往往被认为是社会建立的根基：找到一切事情发生背后的原因。然而其实很多时候，寻找数据间的关联并利用这种关联就足够了。这是我们下一个章节将要讨论的。

---

[1] 计算机象棋的残局的确可以做到完美，但其摧枯拉朽的表现主要还不在于残局。有训练的棋手都能在6个子的情况下不犯错误。这方面的分析和思索，不妨参照一代棋王加里·卡斯帕罗夫（Garry Kasparov）的作品，他亦是对垒“深蓝”的棋王。——译者注

[2] 所有包含不超过6子的残局谱最早是由Unix创造者肯·汤普森发明的，目前的全量残局谱已经可以包含不超过7子的全量局面。——译者注

[3] 特定的大数据企业可以用类似的思路提供实时的指数，例如著名的淘宝消费物价指数（TCPI），其数据来自于淘宝网内消费，可以做到完全实时更新。——译者注

[4] 虽然完全由用户自由添加，标签的形成和组织也有自身的规律。好的标签使用习惯会帮助用户更好管理资源，也会让用户的照片、博客等受到更多关注。相反，胡乱添加标签也会伤害自己。与此同时，标签可以帮助系统提供更好的搜索和推荐服务。关于标签系统的最新研究成果汇总，可以参考张子柯等人2011年在*Journal of Computer Science and Technology*上发表的“Tag-aware recommender systems：A state-of-the-art survey”一文。——译者注

03 更好：不是因果关系，而是相关关系

知道“是什么”就够了，没必要知道“为什么”。在大数据时代，我们不必非得知道现象背后的原因，而是要让数据自己“发声”。

【大数据先锋】

沃尔玛，请把蛋挞与飓风用品摆在一起

FICO,“我们知道你明天会做什么”

美国折扣零售商塔吉特与怀孕预测

UPS与汽车修理预测

大数据预测早产儿病情

幸福感的非线性关系

二手车质量预测

纽约大型沙井盖爆炸预测

林登与亚马逊推荐系统

1997年，24岁的格雷格·林登（Greg Linden）在华盛顿大学就读博士，研究人工智能，闲暇之余，他会在网上卖书。他的网店运营才两年就已经生意兴隆。他回忆说：“我爱卖书和知识，帮助人们找到下一个他们可能会感兴趣的知识点。”他注册的这家网店就是日后大获成功的亚马逊。后来林登被亚马逊聘为软件工程师，以确保网站的正常运行。

亚马逊的技术含量不仅体现在其工作人员上。虽然亚马逊的故事大多数人都耳熟能详，但只有少数人知道它的内容最初是由人工亲自完成的。当时，它聘请了一个由20多名书评家和编辑组成的团队，他们写书评、推荐新书，挑选非常有特色的新书标题放在亚马逊的网页上。这个团队创立了“亚马逊的声音”这个版块，成为当时公司这顶皇冠上的一颗宝石，是其竞争优势的重要来源。《华尔街日报》的一篇文章中热情地

称他们为全美最有影响力的书评家，因为他们使得书籍销量猛增。

杰夫·贝索斯（Jeff Bezos），亚马逊公司的创始人以及总裁，决定尝试一个极富创造力的想法：根据客户个人以前的购物喜好，为其推荐具体的书籍。从一开始，亚马逊已从每一个客户身上捕获了大量的数据。比如说，他们购买了什么书籍？哪些书他们只浏览却没有购买？他们浏览了多久？哪些书是他们一起购买的？

客户的信息数据量非常大，所以亚马逊必须先用传统的方法对其进行处理，通过样本分析找到客户之间的相似性。但这些推荐信息是非常原始的，就如同你在波兰购买一本书，会被东欧其他地区的价格水平搞得晕头转向，或者在买一件婴儿用品时，会被淹没在一堆差不多的婴儿用品中一样。詹姆斯·马库斯（James Marcus）回忆说："推荐信息往往为你提供与你以前购买物品有微小差异的产品，并且循环往复。"詹姆斯·马库斯从1996年到2001年都是亚马逊的书评家，在他的回忆录《亚马逊》（*Amazonia*）里，他说道："那种感觉就像你和一群脑残在一起逛书店。"

格雷格·林登很快就找到了一个解决方案。他意识到，推荐系统实际上并没有必要把顾客与其他顾客进行对比，这样做其实在技术上也比较烦琐。它需要做的是找到产品之间的关联性。1998年，林登和他的同事申请了著名的"item-to-item"协同过滤技术的专利[1]。方法的转变使技术发生了翻天覆地的变化。

因为估算可以提前进行，所以推荐系统快如闪电，而且适用于各种各样的产品。因此，当亚马逊跨界销售除书以外的其他商品时，也可以对电影或烤面包机这些产品进行推荐。由于系统中使用了所有的数据，推荐会更理想。林登回忆道："在组里有句玩笑话，说的是如果系统运作良好，亚马逊应该只推荐你一本书，而这本书就是你将要买的下一本书。"

现在，公司必须决定什么应该出现在网站上。是亚马逊内部书评家写的个人建议和评论，还是由机器生成的个性化推荐和畅销书排行榜？批评者说了什么，或者是顾客的点击意味着什么？从字面上来讲，这是一场人与鼠标的战争。

林登做了一个关于评论家所创造的销售业绩和计算机生成内容所产生的销售业绩的对比测试，结果他发现两者之间的业绩相差甚远。他解

释说，通过数据推荐产品所增加的销售远远超过书评家的贡献。计算机可能不知道为什么喜欢海明威作品的客户会购买菲茨杰拉德的书。但是这似乎并不重要，重要的是销量。最后，编辑们看到了销售额分析，亚马逊也不得不放弃每次的在线评论，最终书评组被解散了。林登回忆说："书评团队被打败、被解散，我感到非常难过。但是，数据没有说谎，人工评论的成本是非常高的。"

如今，据说亚马逊销售额的三分之一都是来自于它的个性化推荐系统。有了它，亚马逊不仅使很多大型书店和音乐唱片商店歇业，而且当地数百个自认为有自己风格的书商也难免受转型之风的影响。事实上，林登的工作彻底改变了电子商务，现在几乎每个人都在使用电子商务。

奈飞公司是一个在线电影租赁公司，它四分之三的新订单都来自推荐系统。在亚马逊的带领下，成千上万的网站可以推荐产品、内容和朋友以及很多相关的信息，但并不知道为什么人们会对这些信息感兴趣。

知道人们为什么对这些信息感兴趣可能是有用的，但这个问题目前并不是很重要。但是，知道"是什么"可以创造点击率，这种洞察力足以重塑很多行业，不仅仅只是电子商务。所有行业中的销售人员早就被告知，他们需要了解是什么让客户做出了选择，要把握客户做决定背后的真正原因，因此专业技能和多年的经验受到高度重视。大数据却显示，还有另外一个在某些方面更有用的方法。亚马逊的推荐系统梳理出了有趣的相关关系，但不知道背后的原因。知道是什么就够了，没必要知道为什么。

关联物，预测的关键

在小数据世界中，相关关系也是有用的，但在大数据的背景下，相关关系大放异彩。通过应用相关关系，我们可以比以前更容易、更快捷、更清楚地分析事物。

大数据洞察

相关关系的核心是量化两个数据值之间的数理关系。相关关系强是指当一个数据值增加时，另一个数据值很有可能也会随之增加。我们已经看到过这种很强的相关关系，比如谷歌流感趋势：在一个特定的地理位置，越多的人通过谷歌搜索特定的词条，该地区就有更多的人患了流感。

相反，相关关系弱就意味着当一个数据值增加时，另一个数据值几乎不会发生变化。[2]例如，我们可以寻找关于个人的鞋码和幸福的相关关系，但会发现它们几乎扯不上什么关系。

相关关系通过识别有用的关联物来帮助我们分析一个现象，而不是通过揭示其内部的运作机制。当然，即使是很强的相关关系也不一定能解释每一种情况，比如两个事物看上去行为相似，但很有可能只是巧合。如实证学家纳西姆·尼古拉斯·塔勒布（Nassim Nicholas Taleb）所言，我们可能只是“被随机性所愚弄”而已。相关关系没有绝对，只有可能性。也就是说，不是亚马逊推荐的每本书都是顾客想买的书。但是，如果相关关系强，一个相关链接成功的概率是很高的。这一点很多人可以证明，他们的书架上有很多书都是因为亚马逊推荐而购买的。

通过给我们找到一个现象的良好的关联物，相关关系可以帮助我们捕捉现在和预测未来。如果A和B经常一起发生，我们只需要注意到B发生了，就可以预测A也发生了。这有助于我们捕捉可能和A一起发生的事情，即使我们不能直接测量或观察到A。更重要的是，它还可以帮助我们预测未来可能发生什么。当然，相关关系是无法预知未来的，他们只能预测可能发生的事情。但是，这已经极其珍贵了。

大数据先锋

沃尔玛，请把蛋挞与飓风用品摆在一起

沃尔玛是世界上最大的零售商，拥有超过200万的员工，销售额约4500亿美元，比大多数国家的国内生产总值还多。在网络带来巨多数据之前，沃尔玛在美国企业中拥有的数据资源应该是最多的。[3]

在20世纪90年代，零售链通过把每一个产品记录为数据而彻底改变了零售行业。沃尔玛可以让供应商监控销售速率、数量以及存货的情况。沃尔玛通过打造透明度来迫使供应商照顾好自己的物流。在许多情况下，沃尔玛不接受产品的“所有权”，除非产品已经开始销售，这样就避免了存货的风险也降低了成本。实际上，沃尔玛运用这些数据使其成为了世界上最大的“寄售店”。

倘若得到正确分析，历史数据能够解释什么呢？零售商与天睿资讯（Teradata）[4]专业的数字统计员一起研究发现了有趣的相关关系。2004年，沃尔玛对历史交易记录这个庞大的数据库进行了观察，这个数据库

记录的不仅包括每一个顾客的购物清单以及消费额，还包括购物篮中的物品、具体购买时间，甚至购买当日的天气。

沃尔玛公司注意到，每当在季节性飓风来临之前，不仅手电筒销售量增加了，而且POP-Tarts蛋挞（美式含糖早餐零食）的销量也增加了。因此，当季节性风暴来临时，沃尔玛会把库存的蛋挞放在靠近飓风用品的位置，以方便行色匆匆的顾客从而增加销量。

过去，总部的人员们需要先有了想法，然后才能收集数据来测试这个想法的可行性。如今，我们有了如此之多的数据和更好的工具，所以要找到相关系变得更快、更容易了。这就意味着我们必须关注：当数据点以数量级方式增长的时候，我们会观察到许多似是而非的相关关系。毕竟我们还处于考察相关关系的初期，所以这一点需要我们高度重视。

在大数据时代来临前很久，相关关系就已经被证明大有用途。这个观点是1888年查尔斯·达尔文的表弟弗朗西斯·高尔顿爵士（Francis Galton）提出的，因为他注意到人的身高和前臂的长度有关系。相关关系背后的数学计算是直接而又有活力的，这是相关关系的本质特征，也是让相关关系成为最广泛应用的统计计量方法的原因。但是在大数据时代之前，相关关系的应用很少。因为数据很少而且收集数据很费时费力，所以统计学家们喜欢找到一个关联物，然后收集与之相关的数据进行相关关系分析来评测这个关联物的优劣。那么，如何寻找这个关联物呢？

除了仅仅依靠相关关系，专家们还会使用一些建立在理论基础上的假想来指导自己选择适当的关联物。这些理论就是一些抽象的观点，关于事物是怎样运作的。然后收集与关联物相关的数据来进行相关关系分析，以证明这个关联物是否真的合适。如果不合适，人们通常会固执地再次尝试，因为担心可能是数据收集的错误，而最终却不得不承认一开始的假想甚至假想建立的基础都是有缺陷和必须修改的。这种对假想的反复试验促进了学科的发展。但是这种发展非常缓慢，因为个人以及团体的偏见会蒙蔽我们的双眼，导致我们在设立假想、应用假想和选择关联物的过程中犯错误。总之，这是一个烦琐的过程，只适用于小数据时代。

在大数据时代，通过建立在人的偏见基础上的关联物监测法已经不再可行，因为数据库太大而且需要考虑的领域太复杂。幸运的是，许多迫使我们选择假想分析法的限制条件也逐渐消失了。我们现在拥有如此

多的数据，这么好的机器计算能力，因而不再需要人工选择一个关联物或者一小部分相似数据来逐一分析了。复杂的机器分析能为我们辨认出谁是最好的代理，就像在谷歌流感趋势中，计算机把检索词条在5亿个数学模型上进行测试之后，准确地找出了哪些是与流感传播最相关的词条。

我们理解世界不再需要建立在假设的基础上，这个假设是指针对现象建立的有关其产生机制和内在机理的假设。因此，我们也不需要建立这样一个假设，关于哪些词条可以表示流感在何时何地传播；我们不需要了解航空公司怎样给机票定价；我们不需要知道沃尔玛的顾客的烹饪喜好。取而代之的是，我们可以对大数据进行相关关系分析，从而知道哪些检索词条是最能显示流感的传播的，飞机票的价格是否会飞涨，哪些食物是飓风期间待在家里的人最想吃的。我们用数据驱动的关于大数据的相关关系分析法，取代了基于假想的易出错的方法。大数据的相关关系分析法更准确、更快，而且不易受偏见的影响。

建立在相关关系分析法基础上的预测是大数据的核心。这种预测发生的频率非常高，以至于我们经常忽略了它的创新性。当然，它的应用会越来越多。

大数据先锋

FICO，“我们知道你明天会做什么”

一个人的信用常被用来预测他/她的个人行为。美国个人消费信用评估公司，也被称为FICO，在20世纪50年代发明了信用分。2011年，FICO提出了“遵从医嘱评分”——它分析一系列的变量来确定这个人是否会按时吃药，包括一些看起来有点怪异的变量。比方说，一个人在某地居住了多久，这个人结婚了没有，他多久换一个工作以及他是否有私家车。这个评分会帮助医疗机构节省开支，因为它们会知道哪些人需要得到它们的用药提醒。有私家车和使用抗生素并没有因果关系，这只是一种相关关系。但是这就足够激发FICO的首席执行官扬言，“我们知道你明天会做什么。”这是他在2011年的投资人大会上说的。

另一个征信机构，益百利（Experian）有一种服务，可以根据个人的信用卡交易记录预测个人的收入情况。通过分析公司拥有的信用卡历史记录数据库和美国国税局的匿名税收数据，益百利能够得出评分结果。

大数据的力量

证明一个人的收入状况要花费10美元左右，但是益百利的预测结果售价不足1美元。

所以有时候，通过代理取得数据信息比自己去操作烦琐的程序要便宜得多。同时还有另一个征信机构出售“支付能力指数”和“可支配支出指数”，这些指数是用来预测一个人的支付能力的。

相关关系的运用更加广泛了。中英人寿保险有限公司（Aviva）是一家大型保险公司，他们想利用信用报告和顾客市场分析数据来作为部分申请人的血液和尿液分析的关联物。这些分析结果被用来找出更有可能患高血压、糖尿病和抑郁症的人。其中用来分析的数据包括好几百种生活方式的数据，比如爱好、常浏览的网站、常看的节目、收入估计等。

Aviva的预测模型是由德勤咨询公司发明的，公司觉得这可以用来发现健康隐患。其他保险公司如英国保诚保险有限公司（Prudentia）和美国国际集团（AIG）都承认它们也在考虑类似的方法。好处就是，申请者不再需要提供血液和尿液样本了，因为大家都不太喜欢做这个事情。

大数据的力量

通过利用相关关系，保险公司可以在每人身上节省125美元，然而这个纯数据分析法只需要花费5美元。

有些人可能会觉得这种方法听起来很恐怖，这些公司似乎可以利用任何网络上的信息。这会让人们下次登陆极限运动网站和坐到电视机前观看幽默情景剧前三思而后行，因为不想因此支付更多的保险费用。让我们在接触任何信息的时候（同时也产生可能被分析的数据）都胆战心惊是一件非常糟糕的事情。但另一方面，这个系统有助于更多的人得到保险，这对于社会和保险公司都是有好处的。

大数据先锋

美国折扣零售商塔吉特与怀孕预测

大数据相关关系分析的极致，非美国折扣零售商塔吉特（Target）莫属了。该公司使用大数据的相关关系分析已经有多年。《纽约时报》的记者查尔斯·杜西格（Charles Duhigg）就在一份报道中阐述了塔吉特公司怎样在完全不和准妈妈对话的前提下预测一个女性会在什么时候怀孕。基本上来说，就是收集一个人可以收集到的所有数据，然后通过相关关系分析得出事情的真实状况。

对于零售商来说，知道一个顾客是否怀孕是非常重要的。因为这是一对夫妻改变消费观念的开始，也是一对夫妻生活的分水岭。他们会开始光顾以前不会去的商店，渐渐对新的品牌建立忠诚。塔吉特公司的市场专员们向分析部求助，看是否有什么办法能够通过一个人的购物方式发现她是否怀孕。

公司的分析团队首先查看了签署婴儿礼物登记簿的女性的消费记录。塔吉特公司注意到，登记簿上的妇女会在怀孕大概第三个月的时候买很多无香乳液。几个月之后，她们会买一些营养品，比如镁、钙、锌。公司最终找出了大概20多种关联物，这些关联物可以给顾客进行“怀孕趋势”评分。这些相关关系甚至使得零售商能够比较准确地预测预产期，这样就能够在孕期的每个阶段给客户寄送相应的优惠券，这才是塔吉特公司的目的。

杜西格在《习惯的力量》（*The Power of Habit*）一书中讲到了接下来发生的事情。一天，一个男人冲进了一家位于明尼阿波利斯市郊的塔吉特商店，要求经理出来见他。他气愤地说：“我女儿还是高中生，你们却给她邮寄婴儿服和婴儿床的优惠券，你们是在鼓励她怀孕吗？”而当几天后，经理打电话向这个男人致歉时，这个男人的语气变得平和起来。他说：“我跟我的女儿谈过了，她的预产期是8月份，是我完全没有意识到这个事情的发生，应该说抱歉的人是我。”

在社会环境下寻找关联物只是大数据分析法采取的一种方式。同样有用的一种方法是，通过找出新种类数据之间的相互联系来解决日常需要。比方说，一种称为预测分析法的方法就被广泛地应用于商业领域，它可以预测事件的发生。这可以指一个能发现可能的流行歌曲的算法系统——音乐界广泛采用这种方法来确保它们看好的歌曲真的会流行；也可以指那些用来防止机器失效和建筑倒塌的方法。现在，在机器、发动机和桥梁等基础设施上放置传感器变得越来越平常了，这些传感器被用来记录散发的热量、振幅、承压和发出的声音等。

一个东西要出故障，不会是瞬间的，而是慢慢地出问题的。通过收集所有的数据，我们可以预先捕捉到事物要出故障的信号，比方说发动机的嗡嗡声、引擎过热都说明它们可能要出故障了。系统把这些异常情况与正常情况进行对比，就会知道什么地方出了毛病。通过尽早地发现异常，系统可以提醒我们在故障之前更换零件或者修复问题。通过找出一个关联物并监控它，我们就能预测未来。

大数据先锋

UPS与汽车修理预测

UPS国际快递公司从2000年就开始使用预测性分析来监测自己全美60000辆车规模的车队，这样就能及时地进行防御性的修理。如果车在路上抛锚损失会非常大，因为那样就需要再派一辆车，会造成延误和再装载的负担，并消耗大量的人力物力，所以以前UPS每两三年就会对车辆的零件进行定时更换。但这种方法不太有效，因为有的零件并没有什么毛病就被换掉了。通过监测车辆的各个部位，UPS如今只需要更换需要更换的零件，从而节省了好几百万美元。有一次，监测系统甚至帮助UPS发现了一个新车的一个零件有问题，因此免除了可能会造成的困扰。

无独有偶，桥梁和建筑物上也被安装了传感器来监测磨损程度。大型化工厂和提炼厂也安装了传感器，因为一旦设备的某一个零件有问题，就只有在更换了零件之后生产才能继续进行。收集和分析数据的花费比出现停产的损失小得多。预测性分析并不能解释故障可能会发生的原因，只会告诉你存在什么问题，也就说它并不能告诉你引擎过热是因为什么，磨损的风扇皮带？没拧紧的螺帽？没有答案。

同样的方法也可以运用在人身上。医院使用医疗设备在病人身上装上各种管线同时得到大量的数据。心电图每秒钟就能产生1000个读数。但是只有部分的数据是被保存使用的，大部分都束之高阁了，即使这些数据都能在一定程度上表现出病人的情况。当与其他病人的数据一起考虑的时候，它们就能显现出哪些治疗方法是有效的。

当收集、存储和分析数据的成本比较高的时候，应该适当地丢弃一些数据。安大略理工大学的卡罗琳·麦格雷戈（Carolyn McGregor）博士和一支研究队伍与IBM一起和很多医院合作，用一个软件来监测处理即时的病人信息，然后把它用于早产儿的病情诊断。系统会监控16个不

同地方的数据，比如心率、呼吸、体温、血压和血氧含量，这些数据可以达到每秒钟1260个数据点之多。

在明显感染症状出现的24小时之前，系统就能监测到早产儿细微的身体变化发出的感染信号。麦格雷戈博士说：“你无法用肉眼看到，但计算机可以看到。”这个系统依赖的是相关关系，而不是因果关系。它告诉你的是会发生什么，而不是为什么发生。这正是这个系统的价值！提早知道病情，医生就能够提早治疗，也能更早地知道某种疗法是否有效，这一切都有利于病人的康复。所以，未来这个系统估计会应用到所有病人身上。这个系统可能不会自己做决定，但是它已经做到了机器能做到的最好，那就是帮助人类做到最好。

惊人的是，麦格雷戈博士的大数据分析法能发现一些与医生的传统看法相违背的相关关系。比如说她发现，稳定的生命体征表明病人发生了严重的感染。这很奇怪，因为医生一般认为恶化的疼痛才是全面感染的征兆。你可以想象，以前医生都是下班的时候看看婴儿床旁边的记录本，觉得病情稳定了，也就下班回家了。只有半夜护士的紧急电话才让他们知道大事不好了，他们的直觉犯了大错误。数据表明，早产儿的稳定不但不是病情好转的标志，反而是暴风雨前的宁静，就像是身体要它的器官做好抵抗困难的准备。但是我们也不太确定，我们不知道具体原因，只是看到了相关关系。这需要海量的数据并且找出隐含的相关性才能发现。但是，大数据挽救了很多生命，这是毫无疑问的。

“是什么”，而不是“为什么”

在小数据时代，相关关系分析和因果分析都不容易，都耗费巨大，都要从建立假设开始。然后我们会进行实验——这个假设要么被证实要么被推翻。但由于两者都始于假设，这些分析就都有受偏见影响的可能，而且极易导致错误。与此同时，用来做相关关系分析的数据很难得到，收集这些数据时也耗资巨大。现今，可用的数据如此之多，也就不存在这些难题了。

当然，还有一种不同的情况也逐渐受到了人们的重视。在小数据时代，由于计算机能力的不足，大部分相关关系分析仅限于寻求线性关系。这个情况随着数据的增加肯定会发生改变。事实上，实际情况远比我们所想象的要复杂。经过复杂的分析，我们能够发现数据的“非线性关系”。

大数据先锋

幸福的非线性关系

多年来，经济学家和政治家一直错误地认为收入水平和幸福感是成正比的。我们从数据图表上可以看到，虽然统计工具呈现的是一种线性关系，但事实上，它们之间存在一种更复杂的动态关系：对于收入水平在1万美元以下的人来说，一旦收入增加，幸福感会随之提升；但对于收入水平在1万美元以上的人来说，幸福感并不会随着收入水平提高而提升。如果能发现这层关系，我们看到的就应该是一条曲线，而不是统计工具分析出来的直线。

这个发现对决策者来说非常重要。如果只看到线性关系的话，那么政策重心应完全放在增加收入上，因为这样才能增加全民的幸福感。而一旦察觉到这种非线性关系，策略的重心就会变成提高低收入人群的收入水平，因为这样明显更划算。

当相关关系变得更复杂时，一切就更混乱了。比如，各地麻疹疫苗接种率的差别与人们在医疗保健上的花费似乎有关联。但是，最近哈佛与麻省理工的联合研究小组发现，这种关联不是简单的线性关系，而是一个复杂的曲线图。和预期相同的是，随着人们在医疗上花费的增多，麻疹疫苗接种率的差别会变小；但令人惊讶的是，当增加到一定程度时，这种差别又会变大。发现这种关系对公共卫生官员来说非常重要，但是普通的线性关系分析师是无法捕捉到这个重要信息的。

如今，专家们正在研发能发现并对比分析非线性关系的必要技术工具[5]。一系列飞速发展的新技术和新软件也从多方面提高了相关关系分析工具发现非因果关系的能力，这就好比立体派画家同时从多个角度来表现女性脸庞的手法。

网络分析行业的出现就是一个最明显的例子。多亏了它，让描绘、测量、计算各节点之间的关系变成了可能，我们可以从Facebook上认识更多的朋友，还可以知道法庭上的一些判决的先例，以及谁给谁打了电话。总之，这些工具为回答非因果关系及经验性的问题提供了新的途径。

在大数据时代，这些新的分析工具和思路为我们提供了一系列新的视野和有用的预测，我们看到了很多以前不曾注意到的联系，还掌握了

以前无法理解的复杂技术和社会动态。但最重要的是，通过去探求“是什么”而不是“为什么”，相关关系帮助我们更好地了解了这个世界。

这听起来似乎有点违背常理。毕竟，人们都希望通过因果关系来了解这个世界。我们也相信，只要仔细观察，就会发现万事万物皆有因缘。了解事情的起因难道不是我们最大的愿望吗？

在哲学界，关于因果关系是否存在的争论已经持续了几个世纪。毕竟，如果凡事皆有因果的话，那么我们就没有决定任何事的自由了。如果说我们做的每一个决定或者每一个想法都是其他事情的结果，而这个结果又是由其他原因导致的，以此循环往复，那么就不存在人的自由意志这一说了——所有的生命轨迹都只是受因果关系的控制了。因此，对于因果关系在世间所扮演的角色，哲学家们争论不休，有时他们认为，这是与自由意志相对立的。当然，关于理论的争辩并不是我们要研究的重点。

大数据洞察

当我们说人类是通过因果关系了解世界时，我们指的是我们在理解和解释世界各种现象时使用的两种基本方法：一种是通过快速、虚幻的因果关系，还有一种就是通过缓慢、有条不紊的因果关系。大数据会改变这两种基本方法在我们认识世界时所扮演的角色。

首先，我们的直接愿望就是了解因果关系。即使无因果联系存在，我们也还是会假定其存在。研究证明，这只是我们的认知方式，与每个人的文化背景、生长环境以及教育水平是无关的。当我们看到两件事情接连发生的时候，我们会习惯性地从因果关系的角度来看待它们。看看下面的三句话：“弗雷德的父母迟到了；供应商快到了；弗雷德生气了。”

我们读到这里时，可能立马就会想到弗雷德生气并不是因为供应商快到了，而是他父母迟到了的缘故。实际上，我们也不知道到底是什么情况。即便如此，我们还是不禁认为这些假设的因果关系是成立的。

普林斯顿大学心理学专家，同时也是2002年诺贝尔经济学奖得主丹尼尔·卡尼曼（Daniel Kahneman）就是用这个例子证明了人有两种思维模式。第一种是不费力的快速思维，通过这种思维方式几秒钟就能得出结果；另一种是比较费力的慢性思维，对于特定的问题，就是需要

考虑到位。

快速思维模式使人们偏向用因果联系来看待周围的一切，即使这种关系并不存在。这是我们对已有的知识和信仰的执著。在古代，这种快速思维模式是很有用的，它能帮助我们在信息量缺乏却必须快速做出决定的危险情况下化险为夷。但是，通常这种因果关系都是并不存在的。

卡尼曼指出，平时生活中，由于惰性，我们很少慢条斯理地思考问题。所以快速思维模式就占据了上风。因此，我们会经常臆想出一些因果关系，最终导致了对世界的错误理解。

父母经常告诉孩子，天冷时不戴帽子和手套就会感冒。然而，事实上，感冒和穿戴之间却没有直接的联系。有时，我们在某个餐馆用餐后生病了的话，我们就会自然而然地觉得这是餐馆食物的问题，以后可能就不再去这家餐馆了。事实上，我们肚子痛也许是因为其他的传染途径，比如和患者握过手之类的。然而，我们的快速思维模式使我们直接将其归于任何我们能在第一时间想起来的因果关系，因此，这经常导致我们做出错误的决定。

与常识相反，经常凭借直觉而来的因果关系并没有帮助我们加深对这个世界的理解。很多时候，这种认知捷径只是给了我们一种自己已经理解的错觉，但实际上，我们因此完全陷入了理解误区之中。就像采样是我们无法处理全部数据时的捷径一样，这种找因果关系的方法也是我们大脑用来避免辛苦思考的捷径。

在小数据时代，很难证明由直觉而来的因果联系是错误的。现在，情况不一样了。将来，大数据之间的相关关系，将经常会用来证明直觉的因果联系是错误的。最终也能表明，统计关系也不蕴含多少真实的因果关系。总之，我们的快速思维模式将会遭受各种各样的现实考验。

令人欣喜的是，为了更好地了解世界，我们会因此更加努力地思考。但是，即使是我们用来发现因果关系的第二种思维方式——慢性思维，也将因为大数据之间的相关关系迎来大的改变。

日常生活中，我们习惯性地用因果关系来考虑事情，所以会认为，因果联系是浅显易寻的。但事实却并非如此。与相关关系不一样，即使用数学这种比较直接的方式，因果联系也很难被轻易证明。我们也不能用标准的等式将因果关系表达清楚。因此，即使我们慢慢思考，想要发

现因果关系也是很困难的。因为我们已经习惯了信息的匮乏，故此亦习惯了在少量数据的基础上进行推理思考，即使大部分时候很多因素都会削弱特定的因果关系。

就拿狂犬疫苗这个例子来说，1885年7月6日，法国化学家路易·巴斯德（Louis Pasteur）接诊了一个9岁的小孩约瑟夫·梅斯特（Joseph Meister），他被带有狂犬病毒的狗咬了。那时，巴斯德刚刚研发出狂犬疫苗，也实验验证过效果了。梅斯特的父母就恳求巴斯德给他们的儿子注射一针。巴斯德做了，梅斯特活了下来。发布会上，巴斯德因为把一个小男孩从死神手中救出而大受褒奖。

但真的是因为他吗？事实证明，一般来说，人被狂犬病狗咬后患上狂犬病的概率只有七分之一。即使巴斯德的疫苗有效，这也只适用于七分之一的案例中。无论如何，就算没有狂犬疫苗，这个小男孩活下来的概率还是有85%。

在这个例子中，大家都认为是注射疫苗救了梅斯特一命。但这里却有两个因果关系值得商榷。第一个是疫苗和狂犬病毒之间的因果关系，第二个就是被带有狂犬病毒的狗咬和患狂犬病之间的因果关系。即便是说疫苗能够医好狂犬病，第二个因果关系也只适用于极少数情况。

不过，科学家已经克服了用实验来证明因果关系的难题。实验是通过是否有诱因这两种情况，分别来观察所产生的结果是不是和真实情况相符，如果相符就说明确实存在因果关系。这个衡量假说的验证情况控制得越严格，你就会发现因果关系越有可能是真实存在的。

因此，与相关关系一样，因果关系被完全证实的可能性几乎是没有的，我们只能说，某两者之间很有可能存在因果关系。但两者之间又有不同，证明因果关系的实验要么不切实际，要么违背社会伦理道德。比方说，我们怎么从5亿词条中找出和流感传播最相关的呢？我们难道真能为了找出被咬和患病之间的因果关系而置成百上千的病人的生命于不顾吗？因为实验会要求把部分病人当成未被咬的“控制组”成员来对待，但是就算给这些病人打了疫苗，我们又能保证万无一失吗？而且就算这些实验可以操作，操作成本也非常的昂贵。

不像因果关系，证明相关关系的实验耗资少，费时也少。与之相比，分析相关关系，我们既有数学方法，也有统计学方法，同时，数字工具也能帮我们准确地找出相关关系。

相关关系分析本身意义重大，同时它也为研究因果关系奠定了基础。通过找出可能相关的事物，我们可以在此基础上进行进一步的因果关系分析，如果存在因果关系的话，我们再进一步找出原因。这种便捷的机制通过严格的实验降低了因果分析的成本。我们也可以从相互联系中找到一些重要的变量，这些变量可以用到验证因果关系的实验中去。

可是，我们必须非常认真。相关关系很有用，不仅仅是因为它能为我们提供新的视角，而且提供的视角都很清晰。而我们一旦把因果关系考虑进来，这些视角就有可能被蒙蔽掉。

例如，Kaggle，一家为所有人提供数据挖掘竞赛平台的公司，举办了关于二手车的质量竞赛。二手车经销商将二手车数据提供给参加比赛的统计学家，统计学家们用这些数据建立一个算法系统来预测经销商拍卖的哪些车有可能出现质量问题。相关关系分析表明，橙色的车有质量问题的可能性只有其他车的一半。

当我们读到这里的时候，不禁也会思考其中的原因。难道是因为橙色车的车主更爱车，所以车被保护得更好吗？或是这种颜色的车子在制造方面更精良些吗？还是因为橙色的车更显眼、出车祸的概率更小，所以转手的时候，各方面的性能保持得更好？

马上，我们就陷入了各种各样谜一样的假设中。若要找出相关关系，我们可以用数学方法，但如果是因果关系的话，这却是行不通的。所以，我们没必要一定要找出相关关系背后的原因，当我们知道了“是什么”的时候，“为什么”其实没那么重要了，否则就会催生一些滑稽的想法。比方说上面提到的例子里，我们是不是应该建议车主把车漆成橙色呢？毕竟，这样就说明车子的质量更过硬啊！

考虑到这些，如果把以确凿数据为基础的相关关系和通过快速思维构想出的因果关系相比的话，前者就更具有说服力。但在越来越多的情况下，快速清晰的相关关系分析甚至比慢速的因果分析更有用和更有效。慢速的因果分析集中体现为通过严格控制的实验来验证的因果关系，而这必然是非常耗时耗力的。

近年来，科学家一直在试图减少这些实验的花费，比如说，通过巧妙地结合相似的调查，做成“类似实验”。这样一来，因果关系的调查成本就降低了，但还是很难与相关关系体现的优越性相抗衡。还有，正如我们之前提到的，在专家进行因果关系的调查时，相关关系分析本来就

会起到帮助的作用。

大数据洞察

在大多数情况下，一旦我们完成了对大数据的相关关系分析，而又不再满足于仅仅知道“是什么”时，我们就会继续向更深层次研究因果关系，找出背后的“为什么”。

因果关系还是有用的，但是它将不再被看成是意义来源的基础。在大数据时代，即使很多情况下，我们依然指望用因果关系来说明我们所发现的相互联系，但是，我们知道因果关系只是一种特殊的相关关系。相反，大数据推动了相关关系分析。相关关系分析通常情况下能取代因果关系起作用，即使不可取代的情况下，它也能指导因果关系起作用。曼哈顿沙井盖（即下水道的修检口）的爆炸就是一个很好的例子。

改变，从操作方式开始

每年，因沙井盖内部失火，纽约每年有很多沙井盖会发生爆炸。重达300磅的沙井盖在轰然塌在地上之前可以冲出几层楼高。这可不是什么好事。

为纽约提供电力支持的联合爱迪生电力公司（Con Edison）每年都会对沙井盖进行常规检查和维修。过去，这完全看运气，如果工作人员检查到的正好是即将爆炸的就最好了，因为沙井盖爆炸威力可不小。2007年，联合爱迪生电力公司向哥伦比亚大学的统计学家求助，希望他们通过对一些历史数据的研究，比如说通过研究以前出现过的问题、基础设施之间的联系，进而预测出可能会出现问题并且需要维修的沙井盖。如此一来，它们就只要把自己的人力物力集中在维修这些沙井盖上。

这是一个复杂的大数据问题。光在纽约，地下电缆就有15万公里，都足够环绕地球三周半了。而曼哈顿有大约51000个沙井盖和服务设施，其中很多设施都是在爱迪生那个时代建成的，而且有二十分之一的电缆在1930年之前就铺好了。尽管1880以来的数据都保存着，却很杂乱，因为从没想过要用来进行数据分析。这些数据都是由会计人员或进行整修的工作人员记录下来的，因为是手记，所以说这些数据杂乱一点也不为过。比如说，常见的“服务设施”代码就有38个之多，而计算机算法需要处理的就是这么混乱的数据：SB，S，S/B，S.B，S?B，S.B.，

SBX，S/BX，SB/X，S/XB，/SBX，S.BX，S＆BX，S?BX，S　BX，S/B/X，S　BOX，SVBX，SERV　BX，SERV-BOX，SERV/BOX，SERVICE BOX……

负责这个项目的统计学家辛西亚·鲁丁（Cynthia Rudin）回忆道：

> 乍看这些数据的时候，我们从未想过能从这些未经处理的数据中找出想要的信息。我打印了一个关于所有电缆的表格。如果把这个表格卷起来的话，除非你在地上拖，不然你绝对提不起它来。而我们需要处理的就是这么多没有处理过的数据。只有理解了这些数据，才能从中淘金，并倾己所有创建一个好的预测模型。

鲁丁和她的同事必须在工作中使用所有的数据，而不能是样本，因为说不定，这成千上万个沙井盖中的某一个就是一个定时炸弹，所以只有使用“样本=总体”的方法才可以。虽然找出因果关系也是不错的，但是这可能需要一个世纪之久，而且还不一定找得对。要完成这项任务，比较好的办法就是，找出它们之间的相关关系。相比“为什么”，她更关心“是什么”。但是她也知道当面对联合爱迪生电力公司高层的时候，她需要证明选择方案的正确性。预测可能是由机器完成的，但是消费者是人类，而人就习惯性地想通过找出原因来理解事物。

鲁丁希望尽快找到整理这些数据的便捷方法。她们将杂乱的数据整理好给机器处理，由此发现了大型沙井盖爆炸的106种预警情况。在布朗克斯（Bronx）的电网测试中，他们对2008年中期之前的数据都进行了分析，并利用这些数据预测了2009年会出现问题的沙井盖。预测效果非常好，在他们列出的前10%的高危沙井盖名单里，有44%的沙井盖都发生了严重的事故。

最终，最重要的因素是这些电缆的使用年限和有没有出现过问题。讽刺的是，这个发现非常有意义，因为联合爱迪生电力公司的高层们可以在此基础上，迅速进行沙井盖事故可能性排序。但是，这些因素看起来会不会太过明显了？

好吧，既是又不是。因为一方面，就像数学家邓肯·沃茨（Duncan Watts）说的，“一旦你知道了结果，一切都很容易。”但是另一方面，我们不能忘记最开始的时候我们可是找出了106种预警情况。如何权衡以及优先修理成千上万个沙井盖中的哪一个，这不是那么容易做出决定的，因为各种各样的因素加入到了这个庞大的数据库中，而且这些数据

记录的方式使得它本来就不适合处理分析。

这个例子说明了数据正在以新的方式帮助我们解决现实生活中的难题。

大数据洞察

我们需要改变我们的操作方式，使用我们能收集到的所有数据，而不仅仅是使用样本。我们不能再把精确性当成重心，我们需要接受混乱和错误的存在。另外，我们应该侧重于分析相关关系，而不再寻求每个预测背后的原因。

大数据，改变人类探索世界的方法

在小数据时代，我们会假想世界是怎么运作的，然后通过收集和分析数据来验证这种假想。在不久的将来，我们会在大数据的指导下探索世界，不再受限于各种假想。我们的研究始于数据，也因为数据我们发现了以前不曾发现的联系。

假想通常来自自然理论或社会科学，它们也是帮助我们解释和预测周遭世界的基础。随着由假想时代到数据时代的过渡，我们也很可能认为我们不再需要理论了。

2008年，《连线》杂志主编克里斯·安德森（Chris Anderson）就指出："数据爆炸使得科学的研究方法都落伍了。"后来，他又在《拍字节时代》（*The Petabyte Age*）的封面故事中讲到，大量的数据从某种程度上意味着"理论的终结"。安德森也表示，用一系列的因果关系来验证各种猜想的传统研究范式已经不实用了，如今它已经被无需理论指导的纯粹的相关关系研究所取代。

为了支撑自己的观点，安德森阐述了量子物理学已变成一门纯理论学科的原因，就是因为实验复杂、耗费多而且不可行。他潜在的观点就是，量子物理学的理论已经脱离实际。[6]他提到了谷歌的搜索引擎和基因排序工程，指出："现在已经是一个有海量数据的时代，应用数学已经取代了其他的所有学科工具。而且只要数据足够，就能说明问题。如果你有一拍字节的数据，只要掌握了这些数据之间的相关关系，一切就都迎刃而解了。"

这篇文章引发了激烈的争论，虽然安德森本人很快就意识到自己的言辞过于激烈了，但是他的观点确实值得深思。安德森的核心思想是，直到目前为止，我们一直都是把理论应用到实践中来分析和理解世界，而如今处在大数据时代，我们不再需要理论了，只要关注数据就足够了。这就意味着所有的普遍规则都不重要了，比方说世界的运作、人类的行为、顾客买什么、东西什么时候会坏等。如今，重要的就是数据分析，它可以揭示一切问题。

大数据洞察

“理论的终结”似乎暗示着，尽管理论仍存在于像物理、化学这样的学科里，但大数据分析不需要成形的概念。这实在荒谬。

大数据是在理论的基础上形成的。比方说，大数据分析就用到了统计和数学理论，有时候也会用到计算机科学理论。是的，这不是关于像地心引力这样特定现象的产生原因的理论，但是无论如何这依然是理论。而且如我们所见，建立在这些理论上的大数据分析模式是实现大数据预测能力的重要因素。事实上，就是因为不受限于传统的思维模式和特定领域里隐含的固有偏见，大数据才能为我们提供如此多新的深刻洞见。

首先就是关于我们怎么收集数据。我们会不会仅仅看数据收集的方便程度来决定呢？或者看数据收集的成本？我们做这些决定的时候就被理论所影响着，而就如达纳·博伊德（Danah Boyd）和凯特·克劳福德（Kate Crawford）说的，我们的选择一定程度上决定了结果。毕竟，谷歌是用检索词来预测流感而不是鞋码。同样，我们在分析数据的时候，也依赖于理论来选择我们使用的工具。最后，我们解读研究结果的时候同样会使用理论。大数据时代绝对不是一个理论消亡的时代，相反地，理论贯穿于大数据分析的方方面面。

作为第一提出问题的人，安德森应该获得掌声——尽管他的答案不怎么样！大数据绝不会叫嚣“理论已死”，但它毫无疑问会从根本上改变我们理解世界的方式。很多旧有的习惯将被颠覆，很多旧有的制度将面临挑战。

大数据时代将要释放出的巨大价值使得我们选择大数据的理念和方法不再是一种权衡，而是通往未来的必然改变。但是在我们到达目的地之前，我们有必要了解怎样才能到达。高科技行业里的很多人认

为是依靠新的工具，从高速芯片到高效软件等。当然，这可以理解为因为他们自己是工具创造者。这些问题固然重要，但不是我们需要考虑的问题。大数据趋势的深层原因，就是海量数据的存在以及越来越多的事物是以数据形式存在的，这也是我们下一章要谈论的内容。

---

[1] 算法思路可参考林登2003年在IEEE Internet Computing上发表的名为“Amazon.com recommendations：item-to-item collaborative filtering”一文。当然，如同谷歌源于PageRank而现在远不仅是PageRank，亚马逊目前的推荐也远远不止基于对象的协同过滤那么简单。举例而言，我所熟悉的百分点通用推荐引擎就包含了数十种常用算法，数千条行业规则和针对用户意图的场景预测模块等。——译者注

[2] 严格地讲，即便没有相关性，另一个数据值也可以大幅变化，只是没有趋势可循罢了。——译者注

[3] 除了利用自有数据外，沃尔玛实验室开始尝试用Facebook好友喜好和Twitter流量与内容分析来实现智能零售。下载一个Facebook上名为shopycat的小应用，就能收到沃尔玛为你创建的礼品清单。——译者注

[4] Teradata的前身是著名的全美现金出纳机公司。——作者注

[5] 2011年，Reshef等人在《科学》上发表了题为“Detecting Novel Associations in Large Datasets”一文，探讨了度量双变量复杂相关行为的新方法。截至目前，该方法还不能处理多变量相关。——译者注

[6] 评价自己不了解的学科，一定要谦虚谨慎，最好是不要做这样的评价。其实，目前的电子产业、纳米加工以及大部分先进的医疗技术都直接来源于量子理论。——译者注

第二部分 大数据时代的商业变革

04 数据化：一切皆可“量化”

大数据发展的核心动力来源于人类测量、记录和分析世界的渴望。信息技术变革随处可见，但是如今的信息技术变革的重点在“T”（技术）上，而不是在“I”（信息）上。现在，我们是时候把聚关灯打向“I”，开始关注信息本身了。

【大数据先锋】

日本先进工业技术研究所的坐姿研究与汽车防盗系统

谷歌的数字图书馆

多效地理定位与UPS的最佳行车路径

Foursquare，让用户在最喜爱的地方“check in”

用手机数据预测疾病传播和城市繁荣

英国对冲基金公司，用微博数据预测股市投资时机

睡眠活动数据库与睡眠模式预测

GPS感应器，判断环境因素对哮喘病的影响

莫里的导航图，大数据的最早实践之一

马修·方丹·莫里（Matthew Fontaine Maury）是一位很有前途的美国海军军官。1839年，在他前往双桅船“合奏号”（Consort）接受一个新任务时，他乘坐的马车突然滑出了车道，瞬间倾倒，把他抛到了空中。他重重地摔到了地上，大腿骨粉碎性骨折，膝盖也脱臼了。当地的医生帮他复位了膝盖关节，但大腿受伤过重，几天后还需要重新手术。直到33岁，他的伤才基本痊愈，但是受伤的腿却留下了残疾，变得有点儿跛，再也无法在海上工作。经过近三年的休养，美国海军把他安排进了办公室，并任命他为图表和仪器厂的负责人。

谁也想不到，这里竟成了他的福地。作为一位年轻的航海家，莫里曾经对船只在水上绕弯儿不走直线而感到十分不解。当他向船长们问及这个问题时，他们回答说，走熟悉的路线比冒险走一条不熟悉而且可能充满危险的路线要好得多。他们认为，海洋是一个不可预知的世界，人随时都可能被意想不到的风浪困住。

但是从他的航行经验来看，莫里知道这并不完全正确。他经历过各种各样的风暴。一次，他听到来自智利瓦尔帕莱索扩展港口的预警，亲眼目睹了当时刮成圆形的风就像钟表一样；但在下午晚些或日落的时候，大风突然结束，静下来变成一阵微风，仿佛有人关了风的开关一样。在另一次远航中，他穿过墨西哥蓝色海域的暖流，感觉就像在大西洋黑黢黢的水墙之间穿行，又好像在密西西比河静止不动的河面上挺进。

当莫里还是一个海军军官学校的学生时，他每次到达一个新的港口，总会向老船长学习经验知识，这些经验知识是代代相传下来的。他从这些老船长那里学到了潮汐、风和洋流的知识，这些都是在军队发的书籍和地图中无法学到的。相反，海军依赖于陈旧的图表，有的都使用了上百年，其中的大部分还有很重大的遗漏和离谱的错误。在他新上任为图表和仪器厂负责人时，他的目标就是解决这些问题。

他清点了库房里的气压计、指南针、六分仪和天文钟。他发现，库房里存放着许多航海书籍、地图和图表；还有塞满了旧日志的发霉木箱，这些都是以前的海军上尉写的航海日志。刚开始的时候，他觉得这些都是垃圾，但当他拍掉被海水浸泡过的书籍上的灰尘，凝视着里面的内容时，莫里突然变得非常激动。

这里有他所需要的信息，例如对特定日期、特定地点的风、水和天气情况的记录。大部分信息都非常有价值。莫里意识到，如果把它们整理到一起，将有可能呈现出一张全新的航海图。这些日志是无章可循的；页面边上尽是奇怪的打油诗和乱七八糟的信手涂鸦，与其说它们是对航海行程的记录，还不如说它们是船员在航海途中无聊的娱乐而已。尽管如此，仍然可以从中提取出有用的数据。莫里和他的20台“计算机”——那些进行数据处理的人，一起把这些破损的航海日志里记录的信息绘制成了表格，这是一项非常繁重的工作。

莫里整合了数据之后，把整个大西洋按经纬度划分成了五块，并按月份标出了温度、风速和风向，因为根据时间的不同这些数据也有所不

同。整合之后，这些数据显示出了有价值的模式，也提供了更有效的航海路线。

有经验的海员有时依靠经验能安全航海，但有时也会陷入危险之中。在从纽约到里约热内卢这条繁忙的航线上，水手们往往倾向于与自然斗争而不是顺应自然。美国船长一直被劝导前往里约热内卢不能通过海峡，因为那样存在很大风险，所以船长会选择在东南方向的航线上航行，再穿过赤道驶向西南方向。而这样一来，航行的距离就相当于穿越大西洋两次。这是很荒谬的，其实直接沿着海峡向南航行就可以了。

为了提高精确度，莫里需要更多的信息，因此他创建了一个标准的表格来记录航海数据，并且要求美国所有的海军舰艇在海上使用，返航后再提交表格。商船也拼命地想得到他的图表，莫里就要求以他们的航海日志作为回报（病毒型社交网络[1]的早期版本）。他宣称："每艘航行在公海上的船舶从此以后都可以被视为一个浮动的天文台，一个科学的殿堂。"为了改进和完善图表，他需要寻求更多的数据（正如谷歌利用网页排名来获得更多的数据）。莫里让船长定期向海里扔掷标有日期、位置、风向以及当时洋流情况的瓶子，然后再来寻找这些瓶子。许多船挂了一面特殊的旗帜，表明它参与了这个信息交流计划。这些旗帜就是出现在一些网站上的友情链接的前身。

通过分析这些数据，莫里知道了一些良好的天然航线，这些航线上的风向和洋流都非常利于航行。他所绘制的图表帮助商人们节省了一大笔钱，因为航海路程减少了三分之一左右。一个船长感激地说："我在得到你的图表之前都是在盲目地航行，你的图表真的指引了我。"有一些顽固的人拒绝使用这个新制的图表，而当他们因为使用旧方法航行到半路出了事故或者花费的航行时间长很多的时候，他们反而帮助证明了莫里系统的实用性。

1855年，莫里的权威著作《关于海洋的物理地理学》（*The Physical Geography of the Sea*）出版，当时他已经绘制了120万数据点了。莫里写道，在这些图表的帮助下，年轻的海员们不用再亲自去探索和总结经验，而能够通过这些图表立即得到来自成千上万名经验丰富的航海家的指导。

他的工作为第一根跨大西洋电报电缆的铺设奠定了基础。同时，在公海上发生了一次灾难性的碰撞事件之后，他马上修改了他的航线分析系统，这个修改后的系统一直沿用至今。他的方法甚至应用到了天文学

领域，1846年当海王星被发现的时候，莫里有了一个好点子，那就是把错把海王星当成一颗恒星时的数据都汇集起来，这样就可以画出海王星的运行轨迹了。

这个土生土长的弗吉尼亚人在美国历史上并不受关注，这也许是因为他在美国内战期间不再为海军效力，而是摇身一变成为了美国联邦政府在英国的间谍。但是多年前，当他前去到欧洲为他绘制的图表寻求国际支持的时候，四个国家授予了他爵士爵位，包括梵蒂冈在内的其他八个国家还颁给了他金牌。即使到今天，美国海军颁布的导航图上仍然有他的名字。

数据，从最不可能的地方提取出来

庞大的数据库有着小数据库所没有的价值，莫里中校是最早发现这一点的人之一。大数据的核心就是挖掘出庞大的数据库独有的价值。更重要的是，他深知只要相关信息能够提取和绘制出来，这些脏乱的航海日志就可以变成有用的数据。通过这样的方式，他重复利用了别人眼里完全没有意义的数据信息。从这个意义上讲，莫里就是数据化的先驱。就像奥伦·埃齐奥尼对Farecast所做的事情一样，用航空业过去的价格信息催生了一个大有赚头的新公司；也像谷歌的工程师所做的一样，通过过去的检索记录预测到了流感的爆发；而莫里则是发挥出了单纯用于记录航行位置的数据的其他用途。

虽然与今天的大数据技术大体类似，但是一想到他是通过人工一笔一画完成的图表绘制，就让人不禁感到惊叹。这说明在远在信息数字化之前，对数据的运用就已经开始了。如今我们经常把“数字化”和“数据化”这两个概念搞混，但是对这两个概念的区分实际上非常重要。我们来看一个更加现代的例子，帮助我们理解数据其实可以从看上去最不可能的东西中提取出来。

大数据先锋

日本先进工业技术研究所的坐姿研究与汽车防盗系统

日本先进工业技术研究所（Japan's Advanced Institute of Industrial Technology）的教授越水重臣（Shigeomi Koshimizu）所做的研究就是关于一个人的坐姿。很少有人会认为一个人的坐姿能表现什么信息，但是它真的可以。当一个人坐着的时候，他的身形、姿势和重量分布都可

以量化和数据化。越水重臣和他的工程师团队通过在汽车座椅下部安装总共360个压力传感器以测量人对椅子施加压力的方式。把人体屁股特征转化成了数据，并且用从0~256这个数值范围对其进行量化，这样就会产生独属于每个乘坐者的精确数据资料。

在这个实验中，这个系统能根据人体对座位的压力差异识别出乘坐者的身份，准确率高达98%。

这个研究并不愚蠢。这项技术可以作为汽车防盗系统安装在汽车上。有了这个系统之后，汽车就能识别出驾驶者是不是车主；如果不是，系统就会要求司机输入密码；如果司机无法准确输入密码，汽车就会自动熄火。把一个人的坐姿转化成数据后，这些数据就孕育出了一些切实可行的服务和一个前景光明的产业。比方说，通过汇集这些数据，我们可以利用事故发生之前的姿势变化情况，分析出坐姿和行驶安全之间的关系。这个系统同样可以在司机疲劳驾驶的时候发出警示或者自动刹车。同时，这个系统不但可以发现车辆被盗，而且可以通过收集到的数据识别出盗贼的身份。

越水重臣教授把一个从不被认为是数据、甚至不被认为和数据沾边的事物转化成了可以用数值来量化的数据模式。同样，莫里中校从看上去没什么用处的事物中提取出了信息，转化成了极其有用的数据。这样创新性的应用创造出了这些信息独特的价值。

数据化，不是数字化

“数据”（data）这个词在拉丁文里是“已知”的意思，也可以理解为“事实”。这是欧几里得的一部经典著作的标题，这本书用已知的或者可由已知推导的知识来解释几何学。如今，数据代表着对某件事物的描述，数据可以记录、分析和重组它。我们还没有合适的词用来形容莫里和越水重臣教授所做的这些转变，所以我们姑且称其为“数据化”吧——这是指一种把现象转变为可制表分析的量化形式的过程。

数据化和数字化大相径庭。数字化指的是把模拟数据转换成用**0**和**1**表示的二进制码，这样电脑就可以处理这些数据了。数字化并不是计算机改革的开始，最初的计算机革命是计算能力的飞跃。我们通过计算机计算过去需要耗费很长时间的项目，比方说导弹弹道表、人口普查结果和天气预报。直到后来才出现了模拟数据和数字化。所以1995年，当美国麻省理工学院媒体实验室的尼古拉斯·尼葛洛庞帝（Nicholas

Negroponte）发表他的标志性著作《数字化生存》（*Being Digital*）的时候，他的主题就是“从原子到比特”。20世纪90年代，我们主要对文本进行数字化。随着过去的几十年里存储能力、处理能力和带宽的提高，我们也能对图像、视频和音乐等类似的内容执行这种转化了。

大数据洞察

今天，技术专家都默认大数据的发展和计算机的变革是同步的。但事实并不是这样的。毫无疑问，是现代信息系统让大数据成为了可能，但是大数据发展的核心动力来源于人类测量、记录和分析世界的渴望。信息技术变革随处可见，但是如今信息技术变革的重点在“T”（技术）上，而不是在“I”（信息）上。现在，我们是时候把聚光灯打向“I”，开始关注信息本身了。

为了得到可量化的信息，我们要知道如何计量；为了数据化量化了的信息，我们要知道怎么记录计量的结果。这需要我们拥有正确的工具。计量和记录的需求也是数据化的前提，而我们在数字化时代来临的几个世纪前就已经奠定好了数据化的基础。

量化一切，数据化的核心

记录信息的能力是原始社会和先进社会的分界线之一。早期文明最古老的抽象工具就是基础的计算以及长度和重量的计量。公元前3000年，信息记录在印度河流域、埃及和美索不达米亚平原地区就有了很大的发展，而日常的计量方法也大有改善。美索不达米亚平原上书写的发展促使了一种记录生产和交易的精确方法的产生，这让早期文明能够计量并记载事实情况，并且为日后所用。计量和记录一起促成了数据的诞生，它们是数据化最早的根基。

计量和记录能够再现人类活动。比如通过记录建筑物的建筑方式和原材料，我们就能再建同样的建筑，或进行实验性的操作，比如通过改变一些方式保存其他部分而建造出新的建筑物，然后再记录这些新建筑物。交易情况一旦得到记录，我们就可以知道一块地丰收时稻谷的产量是多少、需要上缴多少政府税收。计量和记录为预测和计划奠定了基础，虽然这建立在假定明年的收成和今年一样的基础上。有了记录，交易双方才会知道他们赊账的情况，而如果没有这些凭证的支持，欠债的一方则完全可以不用还钱。

几百年来，计量从长度和重量不断扩展到了面积、体积和时间。公元前的最后一个千年，西方的计量方法已经基本准备就绪，但是还是有着比较严重的缺陷。早期文明的计量方法不太适合计算，哪怕是比较简单的计算。比如罗马数字的计算系统就不适合数字计算，因为它没有一个以10为底的记数制或者说是十进制，所以大数目的乘除就算是专家都不知道该怎么算，而简单的乘除对一般人来说也不容易。

大约公元1世纪的时候，印度发明了一种自己的数字系统。它传播到了波斯，并在那里得到改善，而后传入阿拉伯国家，得到了极大的改进。这也就是今天使用的阿拉伯数字的前身。十字军东征给当地人民带来了彻头彻尾的灾难，但同时也把西欧文明带到了地中海东部，而其中最重要的引入就是阿拉伯数字。公元1000年，教皇西尔维斯特二世开始倡导使用阿拉伯数字。12世纪，介绍阿拉伯数字的书籍被翻译成拉丁文，传播到了整个欧洲地区。这也就开启了算术的腾飞。

早在阿拉伯数字传播到欧洲之前，计数板的使用就已经改善了算术。计数板就是在光滑的托盘上放上代币来表示数量，人们通过移动代币到某个区域进行加减。但是，这种计数板有着严重的缺陷，即过大和过小的计算无法同时进行。最主要的缺陷还在于，这些计数板上的数字变化很快，不小心的碰撞或者是摆错一位都会导致完全错误的结果。而且，即便计数板勉强可以进行计算，它也不适合用来记录。因为一旦需要将数字记录在计数板以外的地方，就必须把计数板上的数字转化成罗马数字，这可就费时费力了。[2]

算术赋予了数据新的意义，因为它现在不但可以被记录还可以被分析和再利用。阿拉伯数字从12世纪开始在欧洲出现，而直到16世纪晚期才被广泛采用。到16世纪的时候，数学家们大肆鼓吹他们使用阿拉伯数字计算能比使用计数板快6倍。但最终让阿拉伯数字广为采用的还是复式记账法[3]的出现，它也是数据化的一种工具。

公元前3000年，会计手稿就出现了。但是，记账法在接下来的几百年里发展缓慢，基本上一直保持在记录某地的某个特定交易的阶段。记账人和他的雇主最关心的就是判断某个账户或者自己所从事的行业是否赚钱，而这正是当时的记账手法无法轻易做到的事情。到了14世纪，随着意大利的会计们开始使用两个账本记录交易明细，这种尴尬的境地开始发生改变。这种记账法的优势在于，人们只需要将借贷相加，就可进行制表并得知每个账户的盈亏情况。如此，数据骤然发声了，虽然仅限于读出盈亏情况。

如今，复式记账法通常被看成是会计业和金融业不断发展的成果。事实上，在数据利用的推进过程中，它也是一个里程碑似的存在。它的出现实现了相关账户信息的“分门别类”记录。它建立在一系列记录数据的规则之上，也是最早的信息记录标准化的例子，使得会计们能够读懂彼此的账本。复式记账法可以使查询每个账户的盈亏情况变得简单容易。它会提供交易的记账线索，这样就更容易找到需要的数据。它的设计理念中包含了“纠错”的思想，这也是今天的技术人才们应该学习的。如果一个账本看着不对劲，我们可以查询另一个相对应的账本。

但是，和阿拉伯数字一样，复式记账法也没有立即取得成功。直到200年之后，一个数学家和一个商业家族才让它大受欢迎，他们也改变了数据化的历史。

这个数学家就是方济各会的修士路萨·帕西奥利（Luca Pacioli）。1494年，他出版了一本为普通读者和商人所写的数学教材。这本书大获成功，成为盛行一时的数学教科书。这是第一本全书都使用阿拉伯数字的书籍，因此也促进了阿拉伯数字在欧洲的传播。当然，这本书最大的贡献在于它对复式记账法的详尽论述。接下来的几十年间，这个论述复式记账法的部分被分别译成了6种语言，并且成为几个世纪的通用范本。

而所谓的一个商业家族，就是指美第齐家族——威尼斯商人和艺术资助人。16世纪，这个家族能成为欧洲最有影响力的银行家族，很大一部分要归功于他们使用的一种高级数据记录方法——复式记账法。帕西奥利的著作和美第齐家族的成功奠定了复式记账法成为标准数据记录法的基础，也奠定了阿拉伯数字在此之后不可取代的地位。

伴随着数据记录的发展，人类探索世界的想法一直在膨胀，我们渴望能更精准地记录时间、距离、地点、体积和重量，等等。到了19世纪，随着科学家们发明了新工具来测量和记录电流、气压、温度、声频之类的自然科学现象，科学已经离不开定量化了。那是一个一切事物都需要被测量、划分和记录的时代，人们理解自然的热情甚至高涨到通过分析测量人的颅骨来试图分析人的心智能力。好在，对颅相学这类伪科学的热情最终淡去了，但是人类对于量化一切的热情却始终没有减退。

新工具和开放的思维促进了测量事物和记录数据的繁荣，而现代数据化就诞生于这片沃土之中。数据化的基础已经奠定完好，只是在模拟时代这依然是费时费力的。有时候似乎需要无穷无尽的激情和耐心，或

者说，起码也要有奉献一生的准备，比如16世纪的第谷·布拉赫（Tycho Brahe）就夜夜细心观察天体运动。数据化在模拟时代成功的例子并不多，因为这需要很好的运气——一大串的偶然巧妙地结合在一起。中校莫里就很幸运，他因伤坐进了办公室，但是却在那里发现了珍贵的航海日志，可不是每个人都能这么幸运的。然而，数据化的实现有一点必不可少，那就是要从潜在的数据中挖掘出巨大的价值，然后揭示出新的深刻洞见。

计算机的出现带来了数字测量和存储设备，这样就大大提高了数据化的效率。计算机也使得通过数学分析挖掘出数据更大的价值变成了可能。简而言之，数字化带来了数据化，但是数字化无法取代数据化。数字化是把模拟数据变成计算机可读的数据，和数据化有本质上的不同。

当文字变成数据

数字化和数据化的差异是什么？回答这个问题很容易，我们来看一个两者同时存在并且起作用的领域就可以理解了，这个领域就是书籍。2004年，谷歌发布了一个野心勃勃的计划：它试图把所有版权条例允许的书本内容进行数字化，让世界上所有的人都能通过网络免费阅读这些书籍。为了完成这个伟大的计划，谷歌与全球最大和最著名的图书馆进行了合作，并且还发明了一个能自动翻页的扫描仪，这样对上百万书籍的扫描工作才切实可行且不至于太过昂贵。

刚开始，谷歌所做的是数字化文本，每一页都被扫描然后存入谷歌服务器的一个高分辨率数字图像文件中。书本上的内容变成了网络上的数字文本，所以任何地方的任何人都可以方便地进行查阅了。然而，这还是需要用户要么知道自己要找的内容在哪本书上，要么必须在浩瀚的内容中寻觅自己需要的片段。因为这些数字文本没有被数据化，所以它们不能通过搜索词被查找到，也不能被分析。谷歌所拥有的只是一些图像，这些图像只有依靠人的阅读才能转化为有用的信息。

虽然这是一个现代的、数字化的亚历山大图书馆，比历史上任何一个图书馆都要强大，但谷歌依然希望它能做得更多。谷歌知道，这些信息只有被数据化，它的巨大潜在价值才会被释放出来。因此谷歌使用了能识别数字图像的光学字符识别软件来识别文本的字、词、句和段落，如此一来，书页的数字化图像就转化成了数据化文本。

如今，不仅人类可以使用这些文本信息，计算机也可以处理和分析这些文本数据了。通过检索和查询，我们可以对它进行无穷无尽的文本分析；也可以揭示一个词以及词组第一次出现的时间及其成为流行词的时间，据此发现几百年来人类思维发展和思想传播的轨迹。这种分析支持好几种语言。

大数据先锋

谷歌的数字图书馆

你可以自己试一试。输入网址http：//books.google.com/ngrams，打开Google Ngram Viewer，它利用谷歌所拥有的所有图书作为数据资源，为你提供单词和短语历年使用次数的展示图表。眨眼之间，我们就能发现“causality”（因果关系）这个词在1900年之前的使用频率比“correlation”（相关关系）高；而在1900年之后，情况就与之前相反了。对于作者存在争议的书籍，我们自己也可以进行作品风格鉴定。数据化的实现让抄袭学术作品的行为越来越无处藏身，因此，很多欧洲政客（包括一名德国国防部长）的抄袭行为被曝光，最终不得不引咎辞职。

15世纪中叶，人类发明了印刷机，在这之后大约出版了1.3亿册图书。到2010年为止，也就是谷歌的数字化图书计划实行7年之后，大约有2000万图书被扫描成了数字图书，这几乎相当于人类所有书写文明的15%，这是多么惊人的数字！这诱发了一个新的学术方向——文化组学[4]。“文化组学”是一个计算机专业词汇，指的就是通过文本的定量分析来揭示人类行为和文化发展的趋势。

在一项研究中，哈佛大学的研究员们对几百万册书籍和超过5000亿个单词进行了深入研究，发现这些书中出现过的单词有一半以上在字典中是无法找到的。[5]他们写道，这些因为不够规范而没有录入正规词典中的词汇如此之多，是一个巨大的宝藏。通过系统分析人们如何提及纳粹德国时期的犹太画家马克·夏加尔（Marc Chagall）[6]，他们发现对于思想或是个人的审查和压制会留下“可量化的痕迹”。词语就像是藏于书中而非沉积岩中的化石；信奉“文化组学”的人可以像考古学家一般，挖掘它们所蕴藏的财富。当然，这可能会导致一些可能的偏差，比如图书馆的书籍是不是真实地反映了现实呢？还是反映的只是作者和图书管理员看到的世界？尽管如此，“文化组学”还是会为人们带来很多惊喜的发现。

当文字变成数据，它就大显神通了——人可以用之阅读，机器也可用之分析。但是作为典型的大数据公司，谷歌自然知道收集信息并进行数据化的价值，因为这些数据有非常多的潜在用途。所以，谷歌精明地利用这些数据化了的文本来改进它的机器翻译服务。就像第3章介绍过的一样，这个系统会自动扫描译本，然后找出译语的单词和词组在源语中的对应词和词组是什么。一旦得到答案，系统就可以把翻译看成是一个简单的数学问题，只需要用电脑找出两种语言之间最恰当的对等词和词组。

当然，谷歌并不是第一个梦想在计算机时代唤醒书写文明所蕴含的财富的公司，它也不是第一个吃螃蟹的人。1971年，一个志愿者提出倡议把公共领域的书籍放上网络，制成电子书，方便更多的人阅读，这就是古登堡计划（Project Gutenburg）。这是非常有意义的，但是这个计划没有把书籍数据化，也没有开发出书籍的其他功能；它关注的是阅读，而不是扩充书籍用途。同样地，出版社多年来也一直致力于电子书领域的开发，但是他们都只是把书籍内容作为核心价值，而没有把书籍看作一种数据并纳入自己的商业模式中。因此，他们没有做到把书籍的数据价值挖掘出来，也不允许别人这样做。他们没有看到数据化的需求，也意识不到书籍的数据化潜力。

现在很多公司都在电子书领域激烈地竞争着，亚马逊连同它的Kindle电子书就是这个领域的先驱者。同样在这里，亚马逊和谷歌的发展策略差异显而易见。

亚马逊拥有数据化的书籍，却不曾挖掘书籍数据化之后的附加价值。该公司创始人兼执行总裁贝索斯说服了上百家出版社在Kindle上发布它们的图书，所以Kindle的图书并不是数字图像，人们可以更改字体大小和用黑白以及彩色两种方式看书。这些书籍是数据化了的，不只是数字化。事实上，亚马逊把上百万的新书都数据化了，而谷歌却在费力地数据化很多旧版本的数据。然而，亚马逊把它的眼光聚焦于用来阅读的书籍内容上，而不是分析数据化文本上。当然，有可能它面对了来自传统出版社的压力，后者可能限制了书籍内容的使用方法，毕竟版权在人家手中。谷歌，作为一个喜欢跨界的叛逆的大数据公司，就没有这样的压力了，毕竟谷歌的资源来源于用户点击，而不关出版社什么事。至少现在，可以不失公允地说，亚马逊深谙数字化内容的意义，而谷歌触及了数据化内容的价值。

当方位变成数据

地球本身构成了世界上最基础的信息。但是，历史上它几乎从来没有被量化和数据化过。其实，人和事物的地理定位自然是信息的组成部分，不然我们如何能够吟唱“所谓伊人，在水一方”，但是，这些信息需要转变为数据。

大数据洞察

对地理位置的数据化需要满足一些前提条件。我们需要能精确地测量地球上的每一块地方；我们需要一套标准的标记体系；我们需要收集和记录数据的工具。简而言之，就是地理范围、标准、工具或者说量化、标准化、收集。只有具备了这些，我们才能把位置信息当成数据来存储和分析。

在西方，对位置信息的量化起源于希腊。公元前200年，埃拉托色尼发明了用格网线来划分区域的系统，类似于经纬度法。但是，如同很多古代的好想法一样，它也在历史长河中被慢慢遗忘了。大约1500年之后，也就是公元1400年，托勒密著成的《地理学》（*Geographia*）的复印本从君士坦丁堡传到了佛罗伦萨，那正是文艺复兴和贸易船点燃了对科学和古典知识的热情的时候。著作轰动一时，而书中提到的系统现在仍被用来解决航海导航的难题。从那以后，地图上标上了经纬度和比例尺。这套系统在1570年得到了佛兰德制图师墨卡托的改善，至此海员们就能利用它画出笔直的航线了。

虽然那时就出现了记录地理方位的方法，但却缺乏广为认可的标记标准，使得信息共享依然难以实现。人们急需一套标准的标记系统，就像互联网需要有域名才能正常运行一样。经纬度的标准化是一个漫长的过程。直到1884年，在美国华盛顿召开的国际子午线会议上，25个与会国家中的24个国家一致同意将英格兰格林尼治定为本初子午线和零度经线所穿过的地方（只有自命不凡的法国投了弃权票）。20世纪40年代，墨卡托方位法把世界划分成了60个区域，提高了地理定位的精确性。

这样一来，地理定位信息终于能在标准化的数值范式下进行标记、记录、测量、分析和共享了。但是因为在模拟数据时代，测量和记录地理位置信息耗费很大，人们很少执行。因此，发明能低成本测量地理方位的工具迫在眉睫。到20世纪70年代，进行地理位置定位还只能依靠地标、天文星座、航位推测法和尚显欠缺的无线电定位技术。

1978年见证了一个伟大的转变，当时构成全球定位系统（GPS）的24颗卫星第一次发射成功。无论是汽车上的导航系统还是智能手机，地面上的接收器都能通过计算接收信号所需时间的差异对它们进行三角定位，而这些信号就来自于距离我们20372千米的天空。20世纪80年代，这个系统第一次对民用开放，到90年代才完全投入使用，而同时为了实现商业运用，它的精确度在十年后得到了大幅提升。如今，全球定位系统的地理定位能精确到米，就这样，它实现了自古以来无数航海家、制图家和数学家的梦想。通过与技术手段的融合，全球定位系统能够快速、相对低价地进行地理定位，而且不需要任何专业知识。

定位时时刻刻都可能生成信息。只要愿意，埃拉托色尼或者墨卡托大可以每时每刻都对他们所处的位置进行定位，这谁也管不着。但就算这是可行的，也不现实。同样地，早期的接收器非常复杂和昂贵，适用于潜艇而不是出租车。幸好，改变发生了，这多亏了数字设备中廉价芯片的普及。GPS导航的价格由20世纪90年代的上百美元骤降到了今天的1美元以下。用GPS进行定位一般仅需要几秒钟的时间，它使用的是标准化坐标表示法；所以37°14′06″N　115°48′40″W说明这个人一定是位于内华达州偏远的51号区域（Area 51），一个美军超高安全、超级保密的军事基地——传说那里面关的都是外星人呢！

如今，GPS已经只是众多定位系统中的一种了。中国和欧洲也正在研发新的卫星定位系统来与之抗衡。这些新系统通过对电塔和无线路由器的信号强度进行三角测量来定位地理位置，从而弥补了GPS无法在室内和高楼之间进行定位的缺陷，这也是谷歌、苹果和微软需要自己研发地理定位系统来辅助GPS的原因。谷歌的街景车（Street View Cars）边拍照边收集无线路由器信息；iPhone本身就是一个“移动间谍”，一直在用户不知情的情况下收集位置和无线数据然后传回苹果公司；当然，谷歌的安卓手机和微软的手机操作系统也在收集这一类数据。

除了人以外，我们也可以跟踪事物的地理位置信息。随着汽车装上了无线传感器，地理位置信息的数据化深刻变革了保险的概念。这些数据提供了关于时间、地点和实际行驶路程的详细信息，使保险公司能更好地为车险定价。在英国，车主可以根据他的实际驾驶地点和时间购买汽车保险，而不是只能根据他的年龄、性别和履历来购买年险。这种保险定价法激励投保人产生更好的行为习惯。同时，这改变了保险的基础，从考虑一个群体的平均风险转变为个性化的分析。通过汽车定位每个人的地理方位也改变了一些固定资产投入的模式，比方说公路和其他

基础设施可以让使用这些资源的司机和其他人分担一部分投入。当然，在实现对所有人和事以数据形式保持持续定位之前，这显然还无法实现，但这是我们的发展方向。

大数据先锋

多效地理定位与UPS的最佳行车路径

UPS快递多效地利用了地理定位数据。为了使总部能在车辆出现晚点的时候跟踪到车辆的位置和预防引擎故障，它的货车上装有传感器、无线适配器和GPS。同时，这些设备也方便了公司监督管理员工并优化行车线路。就像莫里的图表是基于过去的航海经验一样，UPS为货车定制的最佳行车路径一定程度上也是根据过去的行车经验总结而来的。

UPS的过程管理总监杰克·莱维斯（Jack Levis）认为这个分析项目效果显著。2011年，UPS的驾驶员们少跑了近4828万公里的路程，节省了300万加仑的燃料并且减少了3万公吨的二氧化碳排放量。系统也设计了尽量少左转的路线，因为左转要求货车在交叉路口穿过去，所以更容易出事故。而且，货车往往需要等待一会儿才能左转，也会更耗油，因此，减少左转使得行车的安全性和效率都得到了大幅提升。

莱维斯说，“预测给我们知识，而知识赋予我们智慧和洞见。”他很确信，有一天，这个系统一定能在用户意识到问题之前预测到并且解决问题。

数据化实时位置信息在人身上的运用最为显著。多年来，无线运营商通过收集和分析这些信息来提升移动互联网的服务水平。不过，这些数据越来越多地被用于其他事情上，第三方也开始利用这些数据来提供新的服务。比方说，一些智能手机的应用程序也不管它本身是否具有定位功能，就收集位置信息；还有一些应用程序就是为了获得用户的位置信息而存在的，比如Foursquare，它让用户在最喜爱的地方“check in”，通过忠诚度计划、酒店推荐和“check in”地点附近的其他推荐而获得好处。

毋庸置疑，收集用户地理位置数据的能力已经变得极其具有价值。从个人层面上来说，根据他所居住的地点和他要去的地方的预测数据，可以为他提供定制广告。而且，这些信息汇集起来可能会揭示事情的发展趋势。[7]比方说，公司可以利用大量的位置数据预测交通情况，你也

许无法想象，这是通过高速公路上的手机而不是汽车的数量和移动速度预测出来的。AirSage每天通过处理来自上百万手机用户的150亿条位置信息，为超过100个美国城市提供实时交通信息。其他两个位置数据服务商Sense Networks和Skyhook使用位置数据揭示城市夜生活最繁荣的地方或者游行队伍聚集了多少人。

不过，位置数据在商业以外的用途或许才是最重要的。麻省理工学院媒体实验室人类动力学[8]实验室主任亚历山大·彭特兰（Alexander "Sandy"Pentland）和他的学生南森·伊格尔（Nathan Eagle）是所谓的"现实挖掘"研究的先驱。**"现实挖掘"**这里指的是通过处理大量来自手机的数据，发现和预测人类行为。在一项研究中，他们通过分析每个人去了哪里、见了谁，成功地区分出了感染了流感的人群，而且在感染者还完全不知道自己已经患病之前就做出了区分。如果出现非常严重的流感疫情，这可以挽救无数人的生命，因为我们会知道应该隔离谁，而且随时都知道去哪里找到他。但是这些数据一旦落入坏人之手，后果将不堪设想，这个问题我们将在后文中继续讨论。

伊格尔是无线数据科技公司Jana的创始人，他使用了来自100多个国家的超过200个无线运营商的手机数据——覆盖了拉丁美洲、非洲、欧洲的大约35亿人口。伊格尔的研究既关注家庭主妇平均每周去几次洗衣店这样的肥皂问题，也试图回答关于疾病如何传播和城市如何繁荣这样的重大问题。在一项研究中，他和同事结合分析了非洲预付费用户的位置信息和他们账户的资费金额，发现资费与收入成正比：越富有的人一次性预付费越多。然而，他们还得出了一个与直觉判断相反的结果，那就是贫民窟不仅仅是永恒不变的贫困中心，还是经济繁荣的跳板。关键就在于，我们要意识到这都是手机所提供的位置信息的间接利用，而和移动通信自身业务没有丝毫关系，但是这些数据最初又是为了更好地开展移动通信而生成的。总之，位置信息一被数据化，新的用途就犹如雨后春笋般涌现出来，而新价值也会随之不断催生。

当沟通变成数据

数据化的另一个前沿更加个人化，直接触摸到了我们的关系、经历和情感。数据化的构思是许多社交网络公司的脊梁。社交网络平台不仅给我们提供了寻找和维持朋友、同事关系的场所，也将我们日常生活的无形元素提取出来，再转化为可作新用途的数据。正因此，Facebook将关系数据化——社交关系在过去一直被视作信息而存在，但从未被正式

界定为数据，直到Facebook“社交图谱”的出现。Twitter通过创新，让人们能轻易记录以及分享他们零散的想法（这些在以前，都会成为遗忘在时光中的碎片），从而使情绪数据化得以实现。LinkedIn将我们过去漫长的经历进行了数据化处理，就像莫里转化旧航海日志那样，把信息转化为对现在和将来的预测：我们可以认识谁，或者哪里存在一份心仪的工作。

然而，数据的使用还远未成熟。就Facebook的情况来说，因为知道太早泄露用户数据的许多新用途会让用户反应过激，所以它精明地选择了忍耐。另外，公司仍然在为其收集的数据数量和类型，包括隐私问题进行商业模式和政策上的调整。目前，它所面对的指责都集中在能采集到什么，而并非它实际用这些数据干了什么。

大数据的力量

Facebook2012年拥有大约10亿用户，他们通过上千亿的朋友关系网相互连接。这个巨大的社交网络覆盖了大约10%的全球总人口。[9]想想这所有的关系和活动在数据化之后都为一家公司所掌控，这些指责和质疑就不能算作空穴来风。

不可否认，其潜在用途非比寻常。一些消费者信贷领域的创业公司正考虑开发以Facebook社交图谱为依据的信用评分。FICO，信用评分系统，利用15个变量来预测单个借贷者是否会偿还一笔债务。但一家获得了高额风险投资的创业公司（很遗憾这里必须匿名）的一项内部研究显示，个人会偿还债务的可能性和其朋友会偿还债务的可能性呈正相关。正应了一句老话：物以类聚，人以群分。因此，Facebook也可以成为下一个FICO。显然，社交媒体上的大量数据也许能形成放飞想象的新型商务基础，其意义远不止表面上我们看到的照片分享、状态上传以及“喜欢”按钮。

同样，Twitter也已经开辟了其数据的新用途。从某种程度上说，2012年超过1.4亿用户每天发送的4亿条微博几乎就和随意的口头零碎差不多。事实上，它们通常就是如此。然而，Twitter公司实现了人们想法、情绪和沟通的数据化，这些都是以前不曾实现的。Twitter与两家公司，DataSift和Gnip达成了一项出售数据访问权限的协议。[10]许多公司对微博做了句法分析，有时还会使用一项叫作情感分析的技术，以获得顾客反馈意见的汇总或对营销活动的效果进行判断。

两家对冲基金，伦敦的英国对冲基金（Derwent Capital）和加利福尼亚的MarketPsych开始分析微博的数据文本，以作为股市投资的信号（他们从未公开自己的商业秘诀，也不知道是倾向于投资势头良好的公司还是做空）。两家公司现在都在向经商者出售信息。就MarketPsych而言，它与Thomson Reuters合作提供了分布在119个国家不低于18864项的独立指数，比如每分钟更新的心情状态，如乐观、忧郁、快乐、害怕、生气，甚至还包括创新、诉讼及冲突情况等。

数据被人类利用的频率远没有被计算机利用得多。以“金融工程师”而闻名的华尔街的数学奇才们，将数据传输到了他们的算法模式当中，来寻找能被有效利用并实现赢利的隐性联系。根据“社交网络分析之父”贝尔纳多·哈柏曼（Bernardo Huberman）[11]的分析，微博中单一主题出现的频率可以用来预测很多事情，比如好莱坞的票房收入。他和一位在惠普实验室工作的同事开发了一个程序，可以用来监听新微博的发布频率，基于此，他们就能预测一部电影的成败，这往往比其他传统评估预测方法还要准确。

这些数据的用途不胜枚举。Twitter微博限制在稀少的140个字符中，但与每条微博联系在一起的元数据是十分丰富的。Twitter的元数据，即“关于信息的信息”，其中包括33个分离的项。虽然一部分信息似乎并没多大用处，比如Twitter用户界面上的“墙纸”或用户用来访问这项服务的软件，但其他的元数据却很有意思，比如他们参与服务所使用的语言、所处的地理位置、关注的人以及粉丝的数量和名字。2011年《科学》杂志上的一项研究显示，来自世界上不同文化背景的人们每天、每周的心情都遵循着相似的模式，这项研究建立在两年多来对84个国家240万人的5.09亿条微博的数据分析上，这在以前是完全无法做到的。情绪真的已经被数据化了。

数据化不仅能将态度和情绪转变为一种可分析的形式，也可能转化人类的行为。这些行为难以跟踪，特别是在较大的社区和其中的子人群环境中。

大数据先锋

微博关联与疫苗接种

来自宾夕法尼亚州立大学的生物学家马塞尔·萨拉特（Marcel Salathé）和软件工程师沙先克·坎都拉斯（Shashank Khandelwal）通过分

析微博发现，人们对于疫苗的态度与他们实际注射预防流感药物的可能呈现出相关性。重要的是，他们利用Twitter用户中谁和谁相关的元数据进行了更进一步的调查，发现未接种疫苗的子人群也可能存在。当然，这项研究的特别之处在于，不同于如谷歌预测流感趋势时利用汇总数据考虑一个地区人口的“平均”健康状况，萨拉特开展的情绪分析实际上揭示了个人的卫生行为。

这些早期的发现预示了数据化将走向何方。和谷歌一样，一些社交网络（如Facebook，Twitter，LinkedIn，Foursquare）坐拥了大型数据的宝藏，一旦这些数据信息得到了深入分析，它们就能轻易获得社会各行各业以及三教九流的几乎所有的动态信息。

世间万物的数据化

只要一点想象，万千事物就能转化为数据形式，并一直带给我们惊喜。IBM获得的“触感技术先导”专利与东京的越水重臣教授对臀部的研究工作具有相同理念。知识产权律师称那是一块触感灵敏的地板，就像一个巨大的智能手机屏幕。其潜在的用途十分广泛。它能分辨出放置其上的物品。它的基本用途就是适时地开灯和开门。然而更重要的是，它能通过一个人的体重、站姿和走路方式确认他的身份。它还能知道某人在摔倒之后是否一直没有站起来。有了它，零售商可以知道商店的人流量。当地板数据化了的时候，它就能滋生无穷无尽的用途。

其实没有听上去那么荒谬。“自我量化”是一项由一群健身迷、医学疯子以及技术狂人发起的运动，通过测量身体的每一个部位和生活中的每一件事来让生活更美好——或者至少用量化的方式来获得新知。目前，自我量化运动规模还很小，但正在日益壮大。

随着智能手机和计算机技术的普及，对个人最重要的生活行为进行数据处理从未如现在这般容易。许多创业公司通过测量人们夜间的脑电波来试图找出他们的睡眠模式。Zeo公司则早已制作出了世界上最大的睡眠活动数据库，揭示了男性与女性睡眠时快速眼动量的差异。Asthmapolis公司将一个感应器绑定到哮喘病人佩戴的呼吸器上，通过GPS定位，再汇总收集起来的位置数据，可以判断环境因素（如接近特定的农作物）对哮喘的影响。Fitbit和Jawbone公司让人们测量他们的体力活动和睡眠。Basis公司用腕带来监测佩戴者的生命体征，包括其心率和皮肤电传导率，以此测试他们所承受的压力。2009年，苹果公司就申请了一项专利，通过音频耳塞收集关于血液氧合、心率和体温的数据。

获取数据正变得比以往任何时候都简单而不受限制。

数据化能帮助我们获取到更多关于人体运作方式的信息。挪威耶维克大学的研究人员和Derawi Biometrics公司联合为智能手机开发了一款应用程序，可以分析人走路时的步伐并将其作为手机解锁的安全系统。同时，佐治亚理工学院的罗伯特·德拉诺（Robert Delano）和布莱恩·派尔思（Brian Parise）开发了一款叫做iTrem的应用程序，用手机内置的测震仪监测人身体的颤动，以应对帕金森和其他神经系统疾病。这个程序给医生和病人都带来了好处；它让患者避免了在医院做昂贵的体检，也让医学专家们能远程监控人们的疾病以及治疗效果。据东京的调查人员说，用智能手机测量震动虽然没有三轴测震仪这种专门的医疗器械那么精确，但也只差了一点，所以完全可以放心使用。这再一次证明，一点点的不精确比完全精确更有效。

在大多数情况下，我们会采集信息并将之存储为数据形式再加以利用。几乎所有领域，任何事情都能这样处理。GreenGoose是一家创业公司，他们销售能放置在物品上的微型运动感应器，用它监测物品的使用次数。比如把它放置在一捆牙线、一个洒水壶或者一盒猫食上，就能数据化牙齿清洁、植物护理以及宠物喂养的信息。很多人对“物联网”有着宗教般的狂热，试图在一切生活中的事物中都植入芯片、传感器和通信模块。这个词听起来好像和互联网亲如姐妹，其实不过是一种典型的数据化手段罢了。

大数据洞察

一旦世界被数据化，就只有你想不到，而没有信息做不到的事情了。莫里通过艰辛的人工分析才揭示了隐藏在数据中的价值，而今天，拥有了数据分析的工具（统计学和算法）以及必需的设备（信息处理器和存储器），我们就可以在更多领域、更快、更大规模地进行数据处理了。在大数据时代，惊喜无处不在！

我们正在进行一个重大的基础设施项目，它在某种程度上与我们过去所做的都不一样，无论是罗马的水渠还是启蒙运动时期的百科全书。它如此的新颖，而我们又深处其中；同时，又因为它是无形的，不像水渠中能触摸到的水，所以我们并未意识到它的存在。这个它，就是无处不在的数据化。像其他的基础设施那样，它会给社会带来根本性的变革。

水渠让城市的发展成为可能，印刷机推进了启蒙运动，报纸为民族国家的兴起奠定了基础。但这些基础设施都侧重于流动——关于水、关于知识。电话和互联网也是如此。相比较而言，数据化代表着人类认识的一个根本性转变。有了大数据的帮助，我们不会再将世界看作是一连串我们认为或是自然或是社会现象的事件，我们会意识到本质上世界是由信息构成的。

整整一个多世纪以来，物理学家们一直宣称情况应该是这样的——并非原子而是信息才是一切的本源。[12]不可否认，这也许听上去无法理解。然而通过数据化，在很多情况下我们就能全面采集和计算有形物质和无形物质的存在，并对其进行处理。

将世界看作信息，看作可以理解的数据的海洋，为我们提供了一个从未有过的审视现实的视角。它是一种可以渗透到所有生活领域的世界观。

大数据洞察

今天，我们生活在一个计算型的社会，因为我们相信世界可以通过数字和数学而获得解释。我们也相信知识可以跨越时空。事实上，我们对书写还存在着一种根深蒂固的敬畏。明天，我们的下一代，一群被“大数据观念”陶冶长大的家伙，会发自肺腑地认为“量化一切”并从中学习对于社会是至关重要的。把各种各样的现实转化为数据，对今天的我们而言也许是新奇而有趣的，但在不久的将来，这将变成如同吃饭睡觉一样与生俱来的能力——这又让我想起了“数据”这个词语的拉丁语原意。

迟早有一天，数据化的影响会使水渠和报纸的影响微乎其微，同时，通过赋予人类数据化世间万物的工具，它也对印刷机和互联网的地位提出了挑战。可是目前，它最主要的用途还是在商业领域。大数据正被用来创造新型价值，这也是下一章的主题。

---

[1]　一种在线社交网络，其典型营销方式是，一旦你加入了，系统会自动从你注册或关联的邮箱中找到联系人，并发信邀请他们加入。很多读者应该都收到过LinkedIn的邀请信，就是一个典型的例子。——译者注

[2] 欧洲人没有接触过东方的算盘，后来证明这是很有利的，因为算盘可能会加剧罗马数字在西方的盛行。——作者注

[3] 所谓复式记账法，是指以资产与权益平衡关系作为记账基础，对于每一项经济业务，都要在两个或两个以上的账户中相互联系进行登记，系统地反映资金运动变化结果的一种记账方法。复式记账的理论依据是会计基本等式，即“资产=负债+所有者权益”。——译者注

[4] 文化组学英文叫做culturomics，是“文化”和“基因组学”两个词的合并。本书中提到的哈佛研究组是文化组学最早的倡导者，其核心人员艾略兹·利波曼·埃顿（Erez Lieberman Aiden）以前是研究基因组学的。——译者注

[5] 出现一次是出现，出现一亿次也是出现。虽然有一半以上的词都是新词怪词，但这些词语出现的频率很低，很多就只出现一两次，而“the”这个单词就出现了数百亿次。所以说，绝大部分单词还是我们认识的。——译者注

[6] 由于是犹太人，他的很多作品都被查封了。——作者注

[7] 通过记录和分析北京市出租车两年的GPS数据，微软亚洲研究院的谢幸及其同事可以向司机提供不同时段的最佳出行路线。——译者注

[8] 人类动力学是一门典型的大数据驱动的定量化学科，它关注人类行为在时间和空间上表现出来的统计规律，以及对这些统计规律的理论解释和在行为预测与控制上的应用。科学出版社2012年出版的《社会动力学》一书有连续6篇论文综述该领域的主要代表方向，可供参考。——译者注

[9] 有部分人在Facebook上拥有多于一个账户。——译者注

[10] 尽管所有微博都是公开的，对“firehose”的访问却需要付费。——作者注

[11] 贝尔纳多·哈柏曼毫无疑问是惠普实验室最有影响力的科学家之一，但是“社会网络之父”这个赞誉还是有些不同寻常。在他11岁的时候，J.A.巴恩斯（J.A.Barnes）已经开始系统研究社交关系，并使用了社

会网络这个概念；他10岁到14岁阶段，正是兰普珀特（A.Rapoport）发展关于社交网络上信息和资源如何扩散、哪些因素导致了社交关系的形成、如何用随机网络和其他数学方法刻画社交网络等一系列研究的关键时间，比哈柏曼更资深、更有影响力的社交网络学者还有很多，譬如林顿·C·弗里曼（Linton Freeman）、马克·格兰诺维特（Mark Granovetter）等。社交网络之父这顶帽子恐怕不应该戴在他的头上，尽管他毫无疑问是非常杰出的科学家。——译者注

[12] 通过阅读Toyabe等人在《自然·物理》上发表的名为“Experimental demonstration of information-to-energy conversion and validation of the generalized Jarzynski equality”一文必会加深读者对该问题的理解。——译者注

05 价值：“取之不尽，用之不竭”的数据创新

数据就像一个神奇的钻石矿，当它的首要价值被发掘后仍能不断给予。它的真实价值就像漂浮在海洋中的冰山，第一眼只能看到冰山的一角，而绝大部分都隐藏在表面之下。

【大数据先锋】

IBM，电动汽车动力与电力供应系统优化预测

Hitwise，通过流量判断消费者喜好

亚马逊，让数据的价值再大一点

移动运营商与数据再利用

谷歌街景与GPS采集

微软与谷歌的拼写检查

谷歌，从大的“噪音”数据中受益

巴诺与NOOK快照

在线教育课程，找到最合适阅读的论坛帖子

Facebook,从66亿到1040亿

DataMarket与InfoChimps，提供免费与付费数据

ReCaptcha与数据再利用

20世纪90年代后期，网络逐渐变得拥堵起来。有人开发了一款名为“Spambots”的垃圾邮件程序软件，向成千上万名用户批量发送广告信息，淹没收件人的电子邮箱。他们会在各种网站上注册，然后在评论部分留下成百上千条广告。网络因此成了一个不守规矩、不受欢迎、不够友善的地方。而且，这种软件似乎打破了网络原有的开放性和易用性模式，要知道，正是这种模式向人们提供了各种便利，比如免费电子邮

件。当特玛捷这一类公司根据“先到先服务”的原则提供演唱会门票网上订票服务时，作弊软件会偷偷摸摸跑到真正排队的人之前，将门票全部买下。

2000年，22岁大学刚毕业的路易斯·冯·安（Luis Von Ahn）提出了解决这个问题的想法：要求注册人提供真实身份证明。他试图找出一些人类容易辨别但对机器来说却很难的东西，最后他想到了一个办法，即在注册过程中显示一些波浪状、辨识度低的字母。人能够在几秒钟内识别并输入正确的文本信息，但电脑却可能会被难倒。雅虎采用了这个方法以后，一夜之间就减轻了垃圾邮件带来的苦恼。冯·安将他的这一创作称为验证码（全称为“全自动区分计算机和人类的图灵测试”）。五年后，每天约有2亿的验证码被用户输入。

这一切给冯·安这位家里经营糖果厂的危地马拉人带来了相当高的知名度，使他能够在取得博士学位后进入卡内基梅隆大学工作，教授计算机科学；也使他在27岁时获得了50万美元的麦克阿瑟基金会“天才奖”。但是，当他意识到每天有这么多人要浪费10秒钟的时间输入这堆恼人的字母，而随后大量的信息被随意地丢弃时，他并没有感到自己很聪明。

于是，他开始寻找能使人的计算能力得到更有效利用的方法。他想到了一个继任者，恰如其分地将其命名为ReCaptcha。和原有随机字母输入不同，人们需要从计算机光学字符识别程序无法识别的文本扫描项目中读出两个单词并输入。其中一个单词其他用户也识别过，从而可以从该用户的输入中判断注册者是人；另一个单词则是有待辨识和解疑的新词。为了保证准确度，系统会将同一个模糊单词发给五个不同的人，直到他们都输入正确后才确定这个单词是对的。在这里，数据的主要用途是证明用户是人，但它也有第二个目的：破译数字化文本中不清楚的单词。ReCaptcha的作用得到了认可，2009年谷歌收购了冯·安的公司，并将这一技术用于图书扫描项目。

大数据的力量

与雇用人所需要花费的成本相比较，它释放出的价值是非常巨大的。每天完成的ReCaptcha超过2亿，按平均每10秒输入一次的话，一天加起来一共是50万个小时，而2012年美国的最低工资是每小时7.25美元。从市场的角度来看，解疑计算机不能识别的单词每天需要花费约350万美元，或者说每年需要花费10亿多美元。冯·安设计的这个系统做

到了这一点，并且，没有花一分钱。

ReCaptcha的故事强调了数据再利用的重要性。随着大数据的出现，数据的价值正在发生变化。

大数据洞察

在数字化时代，数据支持交易的作用被掩盖，数据只是被交易的对象。而在大数据时代，事情再次发生变化。数据的价值从它最基本的用途转变为未来的潜在用途。这一转变意义重大，它影响了企业评估其拥有的数据及访问者的方式，促使甚至是迫使公司改变他们的商业模式，同时也改变了组织看待和使用数据的方式。

信息对于市场交易而言是必不可少的。数据使价格发现成为可能，比如众所周知的一点，它是决定生产数量的信号。一些特殊类型的信息也早已在市场上交易，如书籍、文章、音乐、电影以及金融信息（如股票价格）等。这些在过去的几十年中已经通过个人数据加入数据库。美国的专业数据经纪人，如安客诚（Acxiom）、益百利和艾可飞（Equifax）等，专门负责从数亿名消费者中收集个人信息加入综合档案。随着Facebook、Twitter、LinkedIn、Foursquare等社交平台的出现，我们的人脉关系、想法、喜好和日常生活模式也逐渐被加入到巨大的个人信息库中。

总之，尽管数据长期以来一直是有价值的，但通常只是被视为附属于企业经营核心业务的一部分，或者被归入知识产权或个人信息中相对狭窄的类别。但在大数据时代，所有数据都是有价值的。

这里所说的“所有数据”包含了那些最原始的、看似最平凡的信息单位。想一想工厂机器上热传感器的读数，GPS坐标上的实时数据流，某一辆或者60000辆车的加速度传感器读数和燃料水平。再想想数十亿旧的搜索查询，或者过去数年美国每趟商务航班上每个座位的价格。

但是，直到目前仍然没有一个简单的方法来收集、存储和分析这些数据，这严重限制了提取其潜在价值的机会。在亚当·斯密论述18世纪劳动分工时所引用的著名的大头针制造案例中，监督员需要时刻看管所有工人、进行测量并用羽毛笔在厚纸上记下产出数据，而且测量时间在当时也较难把握，因为可靠的时钟都尚未普及。技术环境的限制使古典经济学家在经济构成的认识上像是戴了一副墨镜，而他们却几乎没有意

识到这一点，就像鱼不知道自己是湿的一样。因此，当他们在考虑生产要素（土地、劳动力和资本）时，信息的作用严重地缺失了。虽然在过去的两个世纪中，数据的采集、存储和使用成本一直在下降，但直到今天也仍然维持在相当昂贵的水平。

我们所处的时代之所以与众不同，是因为数据的收集不再存在固有的局限性。技术已经发展到一定程度，大量信息可以被廉价地捕捉和记录。数据经常会得到被动地收集，人们无须投入太多精力甚至不需要认识这些数据。而且，由于存储成本的大幅下降，保存数据比丢弃数据更加容易。这使得以较低成本获得更多数据的可能性比以往任何时候都大。

大数据的力量

在过去的50年中，数字存储成本大约每两年就削减一半，而存储密度则增加了5000万倍。

在Farecast或谷歌这样的信息公司眼里，数据开始被视为一个新的生产要素，原始材料在数字流水线的一端输入，而处理后的信息则从另一端输出。

大部分数据的直接价值对收集者而言是显而易见的。事实上，数据通常都是为了某个特定的目的而被收集——商店为了会计核算而收集销售数据，工厂为了确保产品符合质量标准而监控输出，网站记录每一个用户点击（即使是鼠标光标的移动）来分析和优化其呈现给访客的内容。数据的基本用途为信息的收集和处理提供了依据。亚马逊同时记录下了客户购买的书籍和他们浏览过的页面，便可以利用这些数据来为客户提供个性化的建议。同样，Facebook跟踪用户的“状态更新”和“喜好”，以确定最佳的广告位从而赚取收入。

不同于物质性的东西，数据的价值不会随着它的使用而减少，而是可以不断地被处理。这就是经济学家所谓的“非竞争性”的好处：个人的使用不会妨碍其他人的使用，而且信息不会像其他物质产品一样随着使用而有所耗损。因此，亚马逊在向其用户，不论是生成这些数据的客户或是其他客户做出建议时，都可以不断地使用过去的交易数据。

大数据洞察

数据的价值并不仅限于特定的用途，它可以为了同一目的而被多次使用，也可以用于其他目的。要了解大数据时代究竟有多少信息对我们有价值，后面这一点尤其重要。

当沃尔玛检查以往的销售数据并发现飓风和蛋挞销售之间存在有利可图的关系时，这种潜力的一部分已经得到实现。这意味着数据的全部价值远远大于其最初的使用价值，也意味着即使首次或之后的每次使用都只带来了少量的价值，但只要数据被多次使用过，企业仍然可以对数据加以有效利用。

数据的“潜在价值”

想知道数据的重复使用对其终极价值有什么意义吗？来看看电动汽车的故事吧。电动汽车能否成功地作为一种交通工具成功普及，其决定因素多如牛毛，但一切都与电池的寿命相关。司机需要能够快速而便捷地为汽车电池充电，电力公司需要确保提供给这些车辆的电力不会影响电网运转。几十年的试验和错误才实现了现有加油站的有效分配，但电动汽车充电站的需求和设置点目前还不得而知。

有趣的是，与其说这是一个基础设施问题，不如说这是一个信息问题，因为大数据是解决方案的重要组成部分。

大数据先锋

IBM，电动汽车动力与电力供应系统优化预测

在2012年进行的一项试验中，IBM曾与加利福尼亚州的太平洋天然气与电气公司以及汽车制造商本田合作，收集了大量信息来回答关于电动汽车应在何时何地获取动力及其对电力供应的影响等基本问题。

基于大量的信息输入，如汽车的电池电量、汽车的位置、一天中的时间以及附近充电站的可用插槽等，IBM开发了一套复杂的预测模型。它将这些数据与电网的电流消耗以及历史功率使用模式相结合。通过分析来自多个数据源的巨大实时数据流和历史数据，能够确定司机为汽车电池充电的最佳时间和地点，并揭示充电站的最佳设置点。最后，系统需要考虑附近充电站的价格差异,即使是天气预报，也要考虑到。例如，如果是晴天，附近的太阳能供电站会充满电，但如果预报未来一周都会下雨，那么太阳能电池板将会被闲置。

系统采用了为某个特定目的而生成的数据，并将其重新用于另一个目的，换言之，数据从其基本用途移动到了二级用途。这使得它随着时间的推移变得更有价值。汽车的电池电量指示器告诉司机应当何时充电，电网的使用数据可以通过设备收集到，从而管理电网的稳定性。这些都是一些基本的用途。这两组数据都可以找到二级用途，即新的价值。它们可以应用于另一个完全不同的目的：确定何时何地充电以及电子汽车服务站的设置点。在此之上，新的辅助信息也将纳入其中，如汽车的位置和电网的历史使用情况。而且，这些数据不只会使用一次，而是随着电子汽车的能耗和电网压力状况的不断更新，一次又一次地为IBM所用。

数据的真实价值就像漂浮在海洋中的冰山，第一眼只能看到冰山一角，而绝大部分则隐藏在表面之下。明白了这一点，那些创新型企业就能够提取其潜在价值并获得潜在的巨大收益。总之，判断数据的价值需要考虑到未来它可能被使用的各种方式，而非仅仅考虑其目前的用途。在我们强调过多次的例子中这一点体现得非常明显：Farecast利用机票销售数据来预测未来的机票价格；谷歌重复使用搜索关键词来监测流感的传播；麦格雷戈博士用婴儿的生命体征来预测传染病的发生；莫里重新利用老船长的日志而发现了洋流。

尽管如此，数据再利用的重要性还没有被企业和社会充分认识到。纽约联合爱迪生公司的高管中很少有谁能够想到，19世纪的电缆信息和工作人员的维修记录可以用来预防未来事故的发生。很多互联网和科技公司甚至直到最近才知道数据再利用具有多大的价值。要解锁这些数据价值，就必须通过新一代统计人员的不懈努力并借助新一代的方法和工具。

用物理学家解释能量的方法或许可以帮助我们理解数据。他们认为物体拥有“储存着的”或“潜在的”能量，只是处于休眠状态，比如压缩了的弹簧或放置在小山顶的小球。这些物体中的能量是隐藏着的、潜在的，直到它们被释放出来。当弹簧被释放或者小球被轻碰而滚下山坡时，这些物体的能量就变成了“动能”，因为它们在移动并对其他物体施力。同理，在基本用途完成后，数据的价值仍然存在，只是处于休眠状态，就像弹簧或小球一样，直到它被二次利用并重新释放它的能量。在大数据时代，我们终于有了这种思维、创造力和工具，来释放数据的隐藏价值。

大数据洞察

最终，数据的价值是其所有可能用途的总和。这些似乎无限的潜在用途就像是选择，这里不是指金融工具意义上的选择，而是实际意义上的选择。这些选择的总和就是数据的价值，即数据的“潜在价值”。

过去，一旦数据的基本用途实现了，我们便认为数据已经达到了它的目的，准备将其删除，让它就此消失。毕竟，数据的首要价值已经得以提取。而在大数据时代，数据就像是一个神奇的钻石矿，在其首要价值被发掘之后仍能不断产生价值。数据的潜在价值有三种最为常见的释放方式：基本再利用、数据集整合和寻找“一份钱两份货”。而数据的折旧值、数据废气和开放数据则是更为独特的方式。

数据创新**1**：数据的再利用

数据创新再利用的一个典型例子是搜索关键词。消费者和搜索引擎之间的瞬时交互形成了一个网站和广告的列表，实现了那一刻的特定功能。乍看起来，这些信息在实现了基本用途之后似乎变得一文不值。但是，以往的查询也可以变得非常有价值。有的公司，如数据代理益百利旗下的网页流量测量公司Hitwise，让客户采集搜索流量来揭示消费者的喜好。通过Hitwise营销人员可以了解到粉红色是否会成为今夏的潮流色，或者黑色是否会回归潮流。谷歌整理了一个版本的搜索词分析，公开供人们查询，并与西班牙第二大银行BBVA合作推出了实时经济指标以及旅游部门的业务预报服务，这些指标都是基于搜索数据得到的。英国央行通过搜索查询房地产的相关信息，更好地了解到了住房价格的升降情况。

大数据先锋

亚马逊，让数据的价值再大一点

未能理解数据再利用重要性的公司以惨痛的代价换来了经验教训。例如，亚马逊早期与AOL达成了一项协议，为AOL电子商务网站提供后台技术服务。在大多数人眼里，这只是一个普通的外包协议，而亚马逊真正的用意在于掌握用户的数据：他们在看什么、买什么。“这些数据可以帮助亚马逊提高它的推荐引擎性能。”亚马逊前首席科学家韦思岸（Andreas Weigend）一语道破。可怜的AOL从来没有意识到这一点，只看到了销售这个基本用途所带来的利益；而聪明的亚马逊却知道如何从二次利用中获利。

再来看另一个例子，谷歌在2007—2010年之间计划在本地搜索列表中加入GOOG—411语音识别服务，但这家搜索巨头并没有自己的语音识别技术，因此急需购买许可。于是，谷歌与该领域的领导者Nuance公司达成合作协议，这家公司因为搭上了这样一个珍贵的客户而感到十分高兴。但Nuance公司在数据方面完全是个十足的笨蛋：合同中没有规定由谁来保存语音翻译记录！于是谷歌自己保存了数据。这些数据在改进技术方面是不可或缺的，谷歌甚至据此从头创建了一个新的语音识别服务系统。当时Nuance公司只考虑到了软件许可的业务交易，而忽视了数据的处理。当认识到自己犯下的错误后，它立即与移动运营商和手机制造商达成其语音识别服务的使用交易，从而进行数据收集。

数据再利用的价值对于那些收集或控制着大型数据集但目前却很少使用的机构来说是个好消息，比如在那些线下运作的传统企业。他们或许正坐在尚未开发的信息喷泉上。有些企业可能已经收集了数据并使用过一次（如果可能的话），且因为存储成本低而将其保存了下来，数据科学家们称这类保存旧信息的计算机为“数据坟墓”。互联网和科技公司在利用海量数据方面走在了最前沿，因为他们仅仅通过在线就能收集大量的信息，分析能力也领先于其他行业。但是，所有的公司都可能会从中获利。麦肯锡的咨询顾问们列举了一家物流公司（名称保密）。这家公司在交付货物的过程中积累了大量产品的全球出货信息。嗅到了这个机会之后，该公司成立了专门的部门，以商业和经济预测的形式出售汇总数据。换言之，它创造了谷歌过去搜索查询业务的一个线下版本。

由于在信息价值链中的特殊位置，有些公司可能会收集到大量的数据，但是他们并不急需使用也并不擅长再次利用这些数据。例如，移动电话运营商收集用户的位置信息来传输电话信号。对于这些公司来说，数据只具有狭窄的技术用途。但是当它被一些发布个性化位置广告服务和促销活动的公司再次利用时，则变得更有价值。有时候，数据的价值并非来自于单个的数据值，而是从数据汇总中体现出来的。因此，AirSage和Sense Networks这些企业会出售诸如人们周五晚上聚集在哪里或者汽车在哪个地段行驶多慢等信息，这种信息集合可以用来确定房地产价值或广告牌的价格。

大数据先锋

移动运营商与数据再利用

如果得到使用正确，即使是最平凡的信息也可以具有特殊的价值。

看看移动运营商吧：他们记录了人们的手机在何时何地连接基站的信息，包括信号的强度。运营商们长期使用这些数据来微调其网络的性能，决定哪里需要添加或者升级基础设施。但这些数据还有很多其他潜在的用途，比如手机制造商可以用它来了解影响信号强度的因素，以改善手机的接收质量。一直以来，处于隐私保护相关法律的限制，移动运营商们并没有用这些数据来谋取利益。但如今，伴随着经济颓势，它们开始逐渐改变立场，认为数据也可以作为其利润的潜在来源。2012年，西班牙电话公司（Telefónica of Spain），一家国际电讯公司，甚至创立了独立公司Telefonica Digital Insights来向零售商和其他买家出售其收集到的匿名用户位置信息。

数据创新2：重组数据

有时，处于休眠状态的数据的价值只能通过与另一个截然不同的数据集结合才能释放出来。用新的方式混合这些数据，我们可以做出很有创意的东西来。一个成功的例子是2011年发表的关于手机是否增加致癌可能性的一项有趣的研究。全球约有60亿部手机，几乎人手一部，因而这个问题是至关重要的。人们做了大量的研究来寻找其中的关联，但都受困于各种障碍：样本量太小、研究时间太短或者是被试自己报告的数据中充满错误。然而，丹麦癌症协会的研究团队基于以往收集的数据想出了一个有趣的方法。

大数据先锋

丹麦癌症协会：手机是否增加致癌率

丹麦拥有1985年手机推出以来所有手机用户的数据库。这项研究分析了1990年至2007年间拥有手机的用户（企业用户和其他社会经济数据不可用的用户除外），共涉及358403人。该国同时记录了所有癌症患者的信息，在那期间共有10729名中枢神经系统肿瘤患者。结合这两个数据集后，研究人员开始寻找两者的关系：手机用户是否比非手机用户显示出较高的癌症发病率？使用手机时间较长的用户是否比时间较短的用户更容易患上癌症？

尽管研究的规模很大，数据却没有出现丝毫混乱或含糊不清。为了满足医疗或商业的目的，两个数据集都采用了严格的质量标准，信息的收集不存在偏差。事实上，数据是在多年前就已经生成的，当时的目的与这项研究毫不相关。最重要的是，这项研究并没有基于任何样本，却

很接近“样本=总体”的准则，即包括了几乎所有癌症患者和移动用户。数据包含了所有的情况，这意味着研究人员掌握了各种亚人群组信息，比如吸烟人群。

最后，研究没有发现使用移动电话和癌症风险增加之间存在任何关系。因此，当2011年10月研究结果在《英国医学杂志》上发布时，并未在媒体中引起任何轰动。但是如果两者之间存在关联的话，它可能马上就会登上世界各地的头版头条，而“重组数据”也可能会随之名声大噪。

随着大数据的出现，数据的总和比部分更有价值。当我们将多个数据集的总和重组在一起时，重组总和本身的价值也比单个总和更大。如今，互联网用户都熟悉基本的混搭式应用，即将两个或多个数据源以一种新颖的方法结合起来。例如，房地产网站Zillow.com将房地产信息和价格添加在美国的社区地图上，同时还聚合了大量的信息，如社区近期的交易和物业规格，以此来预测区域内具体每套住宅的价值。

这个结果极具指导意义，尽管通过视觉展示使得数据更加亲和且非常简单，但采用位置信息并将其置于地图上毕竟不是一个创造性的飞跃。随着大数据的出现，我们可以走得更远，丹麦的癌症研究就为我们提示了更多的可能性。

数据创新3：可扩展数据

促成数据再利用的方法之一是从一开始就设计好它的可扩展性。虽然这不总是可能的，因为人们可能在数据收集后很长时间才意识到这一点，但的确有一些方法可以鼓励相同数据集的多种用途。例如，有些零售商在店内安装了监控摄像头，这样不仅能认出商店扒手，还能跟踪在商店里购物的客户流和他们停留的位置。零售商利用后面的信息可以设计店面的最佳布局并判断营销活动的有效性。在此之前，监控摄像机仅用于安全保卫，是一项纯粹的成本支出，而现在却被视为一项可以增加收入的投资。

大数据先锋

谷歌街景与GPS采集

在收集数据时强调扩展性方面，谷歌毫无疑问是做得最好的公司之一。其备受争议的街景汽车不仅拍摄了房屋和道路的照片，还同时采集

GPS数据，检查地图的信息，甚至还加入了无线网络名称（以及通过开放无线网络的内容，尽管这可能是非法的）。一辆谷歌街景汽车每时每刻都能积累大量的离散数据流。这些数据之所以具有可扩展性，是因为谷歌不仅将其用于基本用途，而且进行了大量的二次使用。例如，GPS数据不仅优化了其地图服务，而且对谷歌自动驾驶汽车的运作功不可没。

收集多个数据流或每个数据流中更多数据点的额外成本往往较低，因此，收集尽可能多的数据并在一开始的时候就考虑到各种潜在的二次用途并使其具有扩展性是非常有意义的。这增加了数据的潜在价值。问题的关键是寻找“一份钱两份货”，即如果以某种方式收集的单一数据集有多种不同的用途，它就具有双重功能。

数据创新4：数据的折旧值

随着数据存储成本的大幅下降，企业拥有了更强的经济动机来保存数据，并再次用于相同或类似的用途。但是，其有效性是有限的。例如，像奈飞和亚马逊这类公司可以利用客户购买的产品、浏览的页面和评论来推荐新的产品，他们可能会年复一年、一遍又一遍地使用这些数据。考虑到这一点，人们可能会认为只要公司不被隐私保护法等法律法规所限制，就应该永远保存这些数字记录，或者至少在经济条件允许的情况下保存这些记录。然而，事实并非如此简单。

随着时间的推移，大多数数据都会失去一部分基本用途。在这种情况下，继续依赖于旧的数据不仅不能增加价值，实际上还会破坏新数据的价值。比如十年前你在亚马逊买了一本书，而现在你可能已经对它完全不感兴趣。如果亚马逊继续用这个数据来向你推荐其他书籍，你就不太可能购买带有这类标题的书籍，甚至会担心该网站之后的推荐是否合理。这些推荐的依据既有旧的过时的信息又有近期仍然有价值的数据，而旧数据的存在破坏了新数据的价值。

于是，亚马逊决定只使用仍有生产价值的数据，这就需要不断地更新数据库并淘汰无用信息。这时面临的挑战就是如何得知哪些数据不再有价值。仅仅依据时间来判断显然不够，[1]因此，亚马逊等公司建立了复杂的模型来帮助自己分离有用和无用的数据。例如，如果客户浏览或购买了一本基于以往购买记录而推荐的书，电子商务公司就认为这项旧的购买记录仍然代表着客户的喜好。这样，他们就能够评价旧数据的有用性，并使模型的“折旧率”更具体。

然而，并非所有的数据都会贬值。有些公司提倡尽可能长时间地保存数据，即使监管部门或公众要求它们短时间内删除或隐匿这些信息。这就解释了为什么一直以来，谷歌都拒绝将互联网协议地址从旧的搜索查询中完全删除（它只是在18个月后删除了最后四位数以隐匿搜索查询）。谷歌希望得到每年的同比数据，如假日购物搜索等。此外，通过了解搜索者的位置，可以帮助改善搜索结果的相关性。例如，很多纽约人都会搜索“火鸡”，但经常会搜索到与“火鸡”无关的关于“土耳其”的网页（英文中“火鸡”与“土耳其”同为turkey）。通过算法可以将他们想要查看的页面放在排名靠前的位置，来方便其他纽约人查找。即使数据用于基本用途的价值会减少，但潜在价值却依然强大。

大数据洞察

潜在价值的概念表明，组织机构应收集尽可能多的使用数据并保存尽可能长的时间。同时也应当与第三方分享数据，前提是要保留所谓的“延展性”权利（专利许可术语）。这样一来，由数据再利用而产生的任何商业价值，原始数据拥有者都能从中分到一杯羹。数据收集者和拥有者无法想象数据再利用的所有可能方式，这一点几乎是不言自明的。

数据创新5：数据废气

数据再利用的方式可以很巧妙、很隐蔽。网络公司可以捕捉到用户在其网站上做的所有事情，然后将每个离散交互当作一个“信号”，作为网站个性化、提高服务或创建全新数字化产品的反馈。两个关于拼写检查的故事给我们提供了一个生动的解释。

大数据先锋

微软与谷歌的拼写检查

在过去的20多年中，微软为其Word软件开发出了一个强大的拼写检查程序，通过与频繁更新的字典正确拼写相比较来对用户键入的字符流进行判断。字典囊括了所有已知词汇，系统将拼写相似但字典中没有的词汇判断为拼写错误，并对其进行纠正。由于需要不断编译和更新字典，微软Word的拼写检查仅适用于最常用的语言，且每年需要花费数百万美元的创建和维护费用。

现在再来看看谷歌是怎么做的吧。可以说，谷歌拥有世界上最完整

的拼写检查器，基本上涵盖了世界上的每一种语言。这个系统一直在不断地完善和增加新的词汇，这是人们每天使用搜索引擎的附加结果。你输错了iPad吗？不要紧，它在那儿呢；Obamacare是什么？哦，明白了。

而且，谷歌几乎是“免费”地获得了这种拼写检查，它依据的是其每天处理的30亿查询中输入搜索框中的错误拼写。一个巧妙的反馈循环可以将用户实际想输入的内容告知系统。当搜索结果页面的顶部显示“你要找的是不是：流行病学”时，用户可以通过点击正确的术语明确地“告诉”谷歌自己需要重新查询的内容。或者，直接在用户访问的页面上显示正确拼写的结果，因为它很可能与正确的拼写高度相关。（这实际上比看上去更有意义，因为随着谷歌拼写检查系统的不断完善，人们即使没有完全精确地输入查询内容也能够获得正确的查询结果。）

谷歌的拼写检查系统显示，那些“不合标准”、“不正确”或“有缺陷”的数据也是非常有用的。有趣的是，谷歌并不是第一个有这种拼写想法的公司。2000年左右，雅虎也看到了从用户输错的查询中创建拼写检查系统的可能性，但只是停留在了想法阶段，并未付诸实践。旧的搜索查询数据就这样被当成了垃圾对待。同样，Infoseek和Alta Vista这两个早期流行的搜索引擎，虽然在那个年代都拥有世界上最全面的错别字数据库，但他们未懂得欣赏其中的价值。在用户不可见的搜索过程中，他们的系统将错别字作为“相关词”进行了处理，但是它的依据是明确告诉系统对与错的字典，而不是鲜活的、有生命的用户交互的总和。

只有谷歌认识到了用户交互的碎屑实际上是金粉，收集在一起就能锻造成一块闪亮的金元宝。谷歌的一名顶级工程师估计，他们的拼写检查器性能比微软至少高出一个数量级（虽然他在采访时承认这并没有进行过可靠计算）。他还嘲笑了“免费”开发的想法——“虽然原材料拼写错误都是免费获得的，但谷歌在系统开发上的花费可能比微软要多得多。”他大笑着说。

这两家公司的不同做法很能说明问题。微软只看到了拼写检查作为文字处理这一个目的的价值，而谷歌却理解了其更深层次的价值。不仅利用错别字开发了世界上最好、最新式的拼写检查器来提高搜索质量，而且将其应用于许多其他服务中，如搜索的“自动完成”功能、Gmail、谷歌文档甚至翻译系统。

一个用来描述人们在网上留下的数字轨迹的艺术词汇出现了，这就

是“数据废气”。它是用户在线交互的副产品，包括浏览了哪些页面、停留了多久、鼠标光标停留的位置、输入了什么信息等。许多公司因此对系统进行了设计，使自己能够得到数据废气并循环利用，以改善现有的服务或开发新服务。毋庸置疑，谷歌是这方面的领导者，它将不断地“从数据中学习”这个原则应用到许多服务中。用户执行的每一个动作都被认为是一个“信号”，谷歌对其进行分析并反馈给系统。

大数据先锋

谷歌，从大的“噪音”数据中受益

例如，谷歌敏锐地注意到，人们经常搜索某个词及其相关词，点击进入后却未能找到想要的信息，于是又返回到搜索页面继续搜索。它知道人们点击的是第1页的第8个链接还是第8页的第1个链接，或者是干脆放弃了所有搜索点击。谷歌不是第一个洞察到这一点的公司，但它利用这一点并取得了非凡的成果。

这些信息是非常有价值的。如果许多用户都点击搜索结果页底部的链接，就表明这个结果更加具有相关性，谷歌的排名算法就会自动地在随后的搜索中将它提到页面中比较靠前的位置（广告也是如此）。一位谷歌的员工说：“我们喜欢从大的‘噪音’数据集中吸取教训。”

数据废气是许多电脑化服务背后的机制，如语音识别、垃圾邮件过滤、翻译等。当用户指出语音识别程序误解了他们的意思时，他们实际上有效地“训练”了这个系统，让它变得更好。

很多企业都开始设计他们的系统，以这种方式收集和使用信息。在Facebook的早期，数据科学家们研究了数据废气的丰富信息，发现人们会采取某种行动（如回帖、点击图标等）的最重要的预测指标就是他们看到了周围的朋友也在这么做。紧接着，Facebook重新设计了它的系统，使每个用户的活动变得可见并广播出去，这为网站的良性循环做出了新的贡献。逐渐地，这个想法从互联网行业传播至可以收集用户反馈的任何公司。

大数据先锋

巴诺与NOOK快照

电子书阅读器捕捉了大量关于文学喜好和阅读人群的数据：读者阅读一页或一节需要多长时间，读者是略读还是直接放弃阅读，读者是否画线强调或者在空白处做了笔记，这些他们都会记录下来。这就将阅读这种长期被视为个人行为的动作转换成了一种共同经验。一旦聚集起来，数据废气可以用量化的方式向出版商和作者展示一些他们可能永远都不会知道的信息，如读者的好恶和阅读模式。这是十分具有商业价值的。电子图书出版公司可以将这些信息卖给出版商，从而帮助改进书籍的内容和结构。例如，巴诺通过分析Nook电子阅读器的数据了解到，人们往往会弃读长篇幅的非小说类书籍。公司从中受到启发，从而推出“Nook快照”，加入了一系列健康和时事等专题的短篇作品。

Udacity、Coursera和EDX等在线教育课程通过跟踪学生的Web交互来寻找最佳的教学方法。班级人数成千上万，产生的数据也十分惊人。教授们现在可以看到，如果大部分学生需要再看一遍课程内容，就可能表明某些地方他们还不太清楚。在斯坦福大学教授安德鲁·恩格（Andrew Ng）讲授的Coursera机器学习课堂上，他注意到约有2000名学生课外作业的答案是错误的，但错误答案居然是相同的。显然，他们都犯了相同的错误，那么是什么呢？

随着一点点的调查，他终于弄清楚了，他们把一个算法里的两个代数方程弄反了。[2]所以如果现在还有其他学生犯同样错误的话，系统不会简单地告诉他们做错了，而是会提示他们去检查算法。这个系统也应用了大数据，通过分析学生看过的每个论坛帖子以及他们是否正确完成课外作业，来预测看过某个帖子之后的学生正确作答的概率，并由此来确定哪些论坛帖子最适合学生阅读。这些都是过去很难得知的，现在却永远地改变了教学方式。

数据废气可以成为公司的巨大竞争优势，也可能成为对手的强大进入壁垒。试想，如果一家新上市的公司设计了一个比当今行业领先者（如亚马逊、谷歌或Facebook等）更优秀的电子商务网站、社交网站或搜索引擎，它也难以同对手竞争，这不仅是因为其经济规模、网络效应或品牌价值不够好，而是因为这些公司收集了来自客户交互的数据废气并纳入到他们的服务中。一个新的在线教育网站有能力与一个已经具备庞大数据库并且由此知道什么最好的对手相抗衡吗？

数据创新6：开放数据

如今，我们很可能认为谷歌和亚马逊等网站是大数据的先驱者，但

事实上，政府才是大规模信息的原始采集者，并且还在与私营企业竞争他们所控制的大量数据。政府与私营企业数据持有人之间的主要区别就是，政府可以强迫人们为他们提供信息，而不必加以说服或支付报酬。因此，政府将继续收集和积累大量的数据。

大数据对于公共部门的适用性同对商业实体是一样的：大部分的数据价值都是潜在的，需要通过创新性的分析来释放。但是，由于政府在获取数据中所处的特殊地位，因此他们在数据使用上往往效率很低。最近有一个想法得到了公认，即提取政府数据价值最好的办法是允许私营部门和社会大众访问。这其实是基于一个原则：国家收集数据时代表的是其公民，因此它也理应提供一个让公民查看的入口，但少数可能会危害到国家安全或他人隐私权的情况除外。

这种想法让“开放政府数据”的倡议响彻全球。开放数据的倡导者主张，政府只是他们所收集信息的托管人，私营部门和社会对数据的利用会比政府更具有创新性。他们呼吁建立专门的官方机构来公布民用和商业数据；而且数据必须以标准的可机读形式展现，以方便人们处理。否则，信息的公开只会是徒有虚名。

2008年1月21日，奥巴马总统在就职的第一天发表了一份总统备忘录，命令美国联邦机构的负责人公布尽可能多的数据，这使开放政府数据的想法取得了极大的进展。“面对怀疑，公开优先。”他这样指示道。这真是一个了不起的声明，特别是与那些作出相反指令的前任们相比。奥巴马的指令促成了data.gov网站的建立，这是美国联邦政府的公开信息资料库。网站从2009年的47个数据集迅速发展起来，到2012年7月三周年时，数据集已达45万个左右，涵盖了172个机构。

即使是在严谨的英国[3]，现在也出现了实质性的转变。英国政府已经颁布相关规定鼓励信息公开，并支持创建由万维网的发明者蒂姆·伯纳斯（Tim Berners-Lee）参与指导的开放式数据中心，这一举措促进了开放数据的新用途并将数据从国家手中解放出来。

欧盟宣布开放数据的举措很快也会遍及整个欧洲。其他国家，如澳大利亚、巴西、智利等也相继出台并实施了开放数据策略。同时，世界各地越来越多的城市和地区也已经加入开放数据的热潮，一些国际组织也是如此，世界银行就公开了数百个之前被限制的关于经济和社会指标方面的数据集。

同时，各种Web开发人员和富有远见的思想家组成了数据团队来最大化开放数据价值，如美国的阳光基金会和英国的开放知识基金会。

大数据先锋

FlyOnTime的航班时间预测

开放数据早期的一个例子，来自美国一个叫FlyOnTime.us的网站。人们可以交互地（从许多其他相互关系中）判断恶劣天气使某一特定机场的航班延迟的可能性有多大。该网站结合了航班信息和互联网免费提供的官方天气预报。它是由开放数据的倡导者开发的，由此来说明美国联邦政府所积累信息的实际使用价值。不仅数据是免费提供的，而且软件代码也是开放源代码，可供人们学习和再次利用。

FlyOnTime.us让数据说话，并且经常语出惊人。人们可以看到，从波士顿到纽约拉瓜迪亚机场的航班因大雾延迟的时间是因雪延迟的两倍。当人们在候机室逗留时，大部分人可能都料想不到这一点，因为他们通常认为雪才是使航班延迟更重要的原因。正是大数据给了人们这种洞察力，只要将交通运输局的历史航班延误数据和美国联邦航空管理局的机场信息，以及美国国家海洋和大气管理局的以往天气报告和国家气象服务的实时状态结合起来，就能揭示这一切。FlyOnTime.us充分体现了一个不收集或控制信息的实体单位是如何像搜索引擎或大零售商一样，能够获取数据并利用其创造价值。

给数据估值

无论是向公众开放还是将其锁在公司的保险库中，数据的价值都难以衡量。来看看2012年5月18日星期五发生的事吧。这一天，28岁的Facebook创始人马克·扎克伯格（Mark Zuckerberg）在位于美国加利福尼亚州门洛帕克市的公司总部，象征性地敲响了纳斯达克的开盘钟。这家宣称全球约每十人中就有一人是其用户的全球最大社交网络公司，开启了其作为上市公司的征程。

和很多新科技股的第一个上市交易日一样，公司股价立即上涨了11%，翻倍增长甚至已经近在眼前。然而就在这一天，怪事发生了。Facebook的股价开始下跌，期间纳斯达克的电脑因出现技术故障曾暂停交易，但仍然于事无补，情况甚至更加恶化。感到异常的股票承销商在摩根士丹利的带领下，不得不支撑股价，最终以略高于发行价收盘。

上市的前一晚，银行对Facebook的定价是每股38美元，总估值1040亿美元（也就是说，大约是波音公司、通用汽车和戴尔电脑的市值之和）。那么事实上Facebook价值多少呢？在2011年供投资者评估公司的审核账目中，Facebook公布的资产为66亿美元，包括计算机硬件、专利和其他实物价值。那么Facebook公司数据库中存储的大量信息，其账面价值是多少呢？零。它根本没有被计入其中，尽管除了数据，Facebook几乎一文不值。

这令人匪夷所思。加特纳市场研究公司（Gartner）的副总裁道格·莱尼（Doug Laney）研究了Facebook在IPO前一段时间内的数据，估算出Facebook在2009年至2011年间收集了2.1万亿条“获利信息”，比如用户的“喜好”、发布的信息和评论等。与其IPO估值相比，这意味着每条信息（将其视为一个离散数据点）都有约4美分的价值。也就是说，每一个Facebook用户的价值约为100美元，因为他们是Facebook所收集信息的提供者。

那么，如何解释Facebook根据会计准则计算出的价值（约63亿美元）和最初的市场估值（1040亿美元）之间会产生如此巨大的差距呢？目前还没有很好的方法能解释这一点。然而人们普遍开始认为，通过查看公司“账面价值”（大部分是有形资产的价值）来确定企业价值的方法，已经不能充分反映公司的真正价值。事实上，账面价值与“市场价值”（即公司被买断时在股票市场上所获的价值）之间的差距在这几十年中一直在不断地扩大。美国参议院甚至在2000年举行了关于将现行财务报告模式现代化的听证会。现行的财务报告模式始于20世纪30年代，当时信息类的企业几乎不存在。现行财务报表模式与现状的差异不仅会影响公司的资产负债表，如果不能正确评估企业的价值，还可能会给企业带来经营风险和市场波动。

公司账面价值和市场价值之间的差额被记为“无形资产”。20世纪80年代中期，无形资产在美国上市公司市值中约占40%，而在2002年，这一数字已经增长为75%。无形资产早期仅包含品牌、人才和战略这些应计入正规金融会计制度的非有形资产部分。但渐渐地，公司所持有和使用的数据也渐渐纳入了无形资产的范畴。

最终，这意味着目前还找不到一个有效的方法来计算数据的价值。Facebook开盘当天，其正规金融资产与其未记录的无形资产之间相差了近1000亿美元，差距几乎是20倍！太可笑了。但是，随着企业找到在资产负债表上记录数据资产价值的方法，这样的差距有一天也必将消除。

人们正在朝着这个方向前进。在美国最大的无线运营商之一工作的一位高级管理人员透露说，数据持有人在认识到数据的巨大价值之后会研究是否在正式的会计条款中将其作为企业的资产。但是，一旦公司的律师得知此事，便会加以阻止。因为把数据计入账面价值可能会使该公司承担法律责任，律师们并不认为这是一个好主意。

同时，投资者也开始注意到数据的潜在价值。拥有数据或能够轻松收集数据的公司，其股价会上涨；而其他不太幸运的公司，就只能眼看着自己的市值缩水。因为这种状况，数据并不要求其价值正式显示在资产负债表中。尽管做起来有困难，市场和投资者还是会给这些无形资产估价，所以Facebook的股价在最初的几个月中一直摇摆不定。但随着会计窘境和责任问题得到缓解，几乎可以肯定数据的价值将显示在企业的资产负债表上，成为一个新的资产类别。

那么，如何给数据估值呢？诚然，计算价值不再是将其基本用途简单地加总。但是如果数据的大部分价值都是潜在的，需要从未知的二次利用提取，那么人们目前尚不清楚应该如何估算它。这个难度类似于在20世纪70年代布莱克-舒尔斯期权定价理论出现前金融衍生品的定价。它也类似于为专利估值，因为随着各种拍卖、交流、私人销售、许可和大量诉讼的出现，一个知识市场正在逐渐兴起。如果不出意外，给数据的潜在价值贴上价格标签会给金融部门带来无限商机。

一个办法是从数据持有人在价值提取上所采取的不同策略入手，最常见的一种可能性就是将数据授权给第三方。在大数据时代，数据持有人倾向于从被提取的数据价值中抽取一定比例作为报酬支付，而不是敲定一个固定的数额。这有点类似于出版商从书籍、音乐或电影的获利中抽取一定比例，作为支付给作者和表演者的特许权使用费；也类似于生物技术行业的知识产权交易，许可人要求从基于他们技术成果的所有后续发明中抽取一定比例的技术使用费。这样一来，各方都会努力使数据再利用的价值达到最大。然而，由于被许可人可能无法提取数据全部的潜在价值，因此数据持有人可能还会同时向其他方授权使用其数据，两边下注以避免损失。因而，“数据滥交”可能会成为一种常态。

一些试图给数据定价的市场如雨后春笋般出现。2008年在冰岛成立的DataMarket向人们提供其他机构（如联合国、世界银行和欧盟统计局等）的免费数据集，靠倒卖商业供应商（如市场研究公司）的数据来获利。另一家新创办的公司InfoChimps，其总部设在得克萨斯州奥斯汀市，希望成为一个信息中间人，供第三方以免费或付费的方式共享他们

的数据。就像易趣给人们提供了一个出售家中搁置不用的物品的平台一样，这些科技创业公司想为任何手中拥有数据的人提供一个出售数据的平台。例如，Import.io鼓励公司授权别人使用自己手中的数据，不然别人也可以从网上免费收集到这些数据。谷歌的前员工吉尔·埃尔巴兹（Gil Elbaz）创办的Factual收集数据，然后制成数据库供需要者使用。

微软也带着它的Windows Azure DataMarket登上了历史舞台。它的目标是专注高质量的数据和监督所提供的产品，其方式和苹果公司监督其应用程序商店中的产品类似。微软假设，一位销售主管在准备Excel表格时可能还需要做一份公司内部数据和来自经济顾问的GDP增长预测的交叉表，那么她只要点击想要购买的数据，后者将瞬间出现在她的电脑屏幕上。

到目前为止，没有人知道估值模型将发挥出怎样的作用。但可以肯定的是，经济正渐渐开始围绕数据形成，很多新玩家可以从中受益，而一些资深玩家则可能会找到令人惊讶的新生机。用硅谷技术专家和科技出版社员工蒂姆·奥莱利（Tim O＆apos；Reilly）的话来说就是，“数据是一个平台”，因为数据是新产品和新商业模式的基石。

大数据洞察

数据价值的关键是看似无限的再利用，即它的潜在价值。收集信息固然至关重要，但还远远不够，因为大部分的数据价值在于它的使用，而不是占有本身。

下一章，我们将考察数据的实际运用和日益崛起的大数据行业。

---

[1] 即使只考虑因时间不同而造成新旧数据价值的不同，也不是一个容易的问题。2010年Koren在《ACM通讯》上题为“Collaborative Filtering with Temporal Dynamics”的文章向我们展示了这一单一特征亦足以巨大地影响推荐的效果。——译者注

[2] 据此猜测是指期望最大化算法，即Expectation Maximization Algorithm。——译者注

[3] 以往大量的政府信息都封锁在英国皇家版权（Crown Copyright）手中，使用起来非常困难和昂贵（比如在线地图公司用到的邮递区号）。——作者注

06 角色定位：数据、技术与思维的三足鼎立

微软以1.1亿美元的价格购买了大数据公司Farecast，而两年后谷歌则以7亿美元的价格购买了给Farecast提供数据的ITA　Software公司。如今，我们正处在大数据时代的早期，思维和技术是最有价值的，但是最终大部分的价值还是必须从数据本身中挖掘。

【大数据先锋】

ITA software与数据授权

VISA＆MasterCard与商户推荐

微软研究中心与再入院率分析

埃森哲与无线传感监测系统

FlightCaster的大数据思维

谷歌与亚马逊，三者兼备

数据中间商，交通数据处理公司Inrix

The-Numbers.com与电影票房预测

苹果，挖出“潜伏”的数据价值

Decide.com与商品价格预测

2011年，西雅图一家叫Decide.com的科技公司推出了一个雄心勃勃的门户网站，它想为无数顾客预测商品的价格。不过它最初计划的业务范围只限于电子产品，包括手机、平板电视、数码相机等。公司的计算机会收集电子商务网站上所有电子产品的价格数据和产品信息。

网络产品的价格受一系列因素的影响全天都在不断更新，所以公司收集的价格数据必须是即时的。这不仅是一个“大数据”问题，还是一个“大文本”问题，因为系统必须进行数据分析，才会知道一个产品是不

是下架了或者是不是有新产品要发布了，这些都是用户想知道的信息而且都会影响产品价格。

经过一年的时间，Decide.com分析了近400万产品的超过250亿条价格信息。它发现了一些过去人们无法意识到的怪异现象，比如在新产品发布的时候，旧一代的产品可能会经历一个短暂的价格上浮。大部分人都习惯性地认为旧产品更便宜，所以会选择买旧产品，其实这取决于你什么时候购买，不然有可能你付出的金钱比购买新产品还要多。因为电子商务网站都开始使用自动定价系统，所以Decide.com能够发现不正常、不合理的价格高峰，然后告知用户何时才是购买电子产品的最佳时机。

大数据的力量

根据公司内部分析显示，它的预测准确率可以达到77%，平均可以帮助每个顾客在购买一个产品时节省100美元。

表面上，Decide.com就像众多前途光明的科技公司一样，在创新地使用数据、赚取利润。但是事实上，让Decide.com异军突起的不是数据，不是技术，而是思维观念。Decide.com使用的数据都来自电子商务网站和互联网，这是公开的数据，每个人都可以利用。技术上，公司也并没有无可替代的技术人才。所以，虽然数据和技术也是不可或缺的，但是真正使得该公司取得成功的是他们拥有大数据的思维观念。它先人一步地挖掘出了数据的潜在价值。Decide.com和Farecast之间似乎有着相通性，如果你知道它们都是华盛顿大学奥伦·埃齐奥尼先生的杰作，你就知道原因了。

上一章中，我们讨论了怎样通过创新用途，挖掘出数据新的价值，主要是指我们所说的潜在价值。如今，我们的重点转移到了使用数据的公司和它们如何融入大数据价值链中。我们将讨论这对公司、个人的事业和生活意味着什么。

大数据价值链的3大构成

根据所提供价值的不同来源，分别出现了三种大数据公司。这三种来源是指：数据本身、技能与思维。

第一种是基于数据本身的公司。这些公司拥有大量数据或者至少

可以收集到大量数据，却不一定有从数据中提取价值或者用数据催生创新思想的技能。最好的例子就是Twitter，它拥有海量数据这一点是毫无疑问的，但是它的数据都通过两个独立的公司授权给别人使用。

第二种是基于技能的公司。它们通常是咨询公司、技术供应商或者分析公司。它们掌握了专业技能但并不一定拥有数据或提出数据创新性用途的才能。比方说，沃尔玛和Pop-Tarts这两个零售商就是借助天睿公司（Teradata）的分析来获得营销点子，天睿就是一家大数据分析公司。

第三种是基于思维的公司。皮特·华登（Pete Warden），Jetpac的联合创始人，就是通过想法获得价值的一个例子。Jetpac通过用户分享到网上的旅行照片来为人们推荐下次旅行的目的地。对于某些公司来说，数据和技能并不是成功的关键。让这些公司脱颖而出的是其创始人和员工的创新思维，他们有怎样挖掘数据的新价值的独特想法。

大数据洞察

到目前为止，前两种因素一直备受关注，因为在现今世界，技能依然欠缺，而数据则非常之多。近年来，一种新的职业出现了，那就是“数据科学家”。数据科学家是统计学家、软件程序员、图形设计师与作家的结合体。与通过显微镜发现事物不同，数据科学家通过探寻数据库来得到新的发现。全球知名咨询管理公司麦肯锡，就曾极端地预测数据科学家是当今和未来稀缺的资源。如今的数据科学家们也喜欢用这个预测来提升自己的地位和工资水平。

同时，谷歌的首席经济学家哈尔·范里安（Hal Varian）认为统计学家是世界上最棒的职业，他的这种说法非常著名。“如果你想成功，你不应该成为一个普通的、可被随意替代的人，你应该成为稀缺的、不可替代的那类人，”他还说，“数据非常之多而且具有战略重要性，但是真正缺少的是从数据中提取价值的能力。这也就是为什么统计学家、数据库管理者和掌握机器理论的人是真正了不起的人。”

但是，过分强调技术和技能而忽视数据本身的重要性也是不可取的。随着计算机行业的发展，人力技术的落后会被慢慢地克服，而范里安所赞赏的技能将会变成十分普通的事情。认为当今世界数据非常之多，所以收集数据很简单而且数据价值并不高的想法是绝对错误的——数据才是最核心的部分。要知道原因，就必须考虑到大数据价值链的

各个部分，以及它们会如何发展变化。

大数据掌控公司

大数据最值钱的部分就是它自身，所以最先考虑数据拥有者才是明智的。他们可能不是第一手收集数据的人，但是他们能接触到数据、有权使用数据或者将数据授权给渴望挖掘数据价值的人。

大数据先锋

ITA Software与数据授权

四大机票预订系统之一的ITA Software[1]就为Farecast提供预测机票价格所需要的数据，而它自身并不进行这种数据分析。为什么呢？因为商业定位不一样，毕竟出售机票已经很不容易了，所以ITA并不考虑这些数据的额外利用。因此，两家公司的核心竞争力也会不同。当然，还有就是ITA并没有这种创新想法，如果它能像Farecast一样利用数据，那么就需要向奥伦·埃齐奥尼先生购买专利使用权了。

当然，它在大数据价值链上所处的位置也决定了它不会这样去使用数据。“ITA会尽量避免用任何数据来暴露航空公司的利润问题。”ITA的创始人之一也是前CTO卡尔·德马肯（Carl de Marcken）如是说。他还说，“ITA能够得到这些数据而且必须拥有这些数据，因为它们是ITA在提供服务时必须具备的。”但是，ITA有意与这些数据保持一定的距离，所以自己不使用而是授权别人使用。结果不难预见，ITA只从Farecast那里分得了小小的一杯羹。Farecast得到了数据大部分的间接价值，它把其中一部分价值以更便宜的机票的形式转移给了它的用户，而把这种价值带来的利润分给了它的股东以及员工。Farecast通过广告、佣金，当然最后通过出售公司本身获取利润。

有的公司精明地把自己放在了这个信息链的核心，这样它们就能扩大规模、挖掘数据的价值。信用卡行业的情况就符合这一点。多年来，防范信用诈骗的高成本使得许多中小银行都不愿意发行自己的信用卡；而是由大型金融机构发行，因为只有它们才能大规模地投入人力物力发展防范技术。美国第一资本银行和美国银行这样的大型金融机构就承担了这个工作。但是现在小银行后悔了，因为没有自己发行的信用卡，它们就无从得知客户的消费模式，从而不能为客户提供定制化服务。

大数据先锋

VISA＆MasterCard与商户推荐

相对地，像VISA和MasterCard这样的信用卡发行商和其他大银行就站在了信息价值链最好的位置上。通过为小银行和商家提供服务，它们能够从自己的服务网获取更多的交易信息和顾客的消费信息。它们的商业模式从单纯的处理支付行为转变成了收集数据。接下来的问题就是，如何使用收集到的数据。

就像ITA一样，MasterCard也可以把这些数据授权给第三方使用，但是它更倾向于自己分析、挖掘数据的价值。一个称为MasterCard Advisors的部门收集和分析了来自210个国家的15亿信用卡用户的650亿条交易记录，用来预测商业发展和客户的消费趋势。然后，它把这些分析结果卖给其他公司。它发现，如果一个人在下午四点左右给汽车加油的话，他很可能在接下来的一个小时内要去购物或者去餐馆吃饭，而这一个小时的花费大概在35~50美元之间。商家可能正需要这样的信息，因为这样它们就能在这个时间段的加油小票背面附上加油站附近商店的优惠券。

处于这个数据链的中心，MasterCard占据了收集数据和挖掘数据价值的黄金位置。我们可以想象，未来的信用卡公司不会再对交易收取佣金，而是免费提供支付服务。作为回报，它们会获得更多的数据，而对这些数据进行复杂的分析之后，它们又可以卖掉分析结果以取得利润。

大数据技术公司

第二种类型就是拥有技术和专业技能的公司。MasterCard选择了自己分析，有的公司选择在两个类型之间游移，但是还有一部分公司会选择发展专业技能。比方说，埃森哲咨询公司就与各行各业的公司合作应用高级无线感应技术来收集数据，然后对这些数据进行分析。

大数据的力量

2005年，在埃森哲与密苏里州圣路易斯市共同合作的一个实验项目中，它给20辆公交车安装了无线传感器来监测车辆引擎的工作情况。这些数据被用来预测公交车什么时候会抛锚以及维修的最佳时机。研究促使车辆更换零件的周期从30万或者40万公里变成了50万公里，仅这一项

研究结果就帮助该城市节省了60万美元。在这里，获益的不是埃森哲，而是圣路易斯市。

在医学数据领域，我们可以看到一个关于技术公司如何能提供有效服务的很好的例子。位于华盛顿州的华盛顿中心医院与微软研究中心合作分析了多年来的匿名医疗记录，涉及患者人口统计资料、检查、诊断、治疗资料，等等。这项研究是为了减少感染率和再入院率，因为这两项所耗费的费用是医疗卫生领域最大的一部分，所以任何可以减少哪怕是很小比例的方法都意味着节省巨大的开支。

这项研究发现了很多惊人的相关关系：在一系列情况下，一个出院了的病人会在一个月之内再次入院。有一些情况是众所周知但还没有找到好的解决办法的，比方说，一个患充血性心力衰竭的病人就很有可能再次入院，因为这是非常难医治的病。但是研究也发现了一个出人意料的重要因素，那就是病人的心理状况。如果对病人最初的诊断中有类似“压抑”这种暗示心理疾病的词的话，病人再度入院的可能性大很多。

虽然这种相关关系对于建立特定的因果关系并无帮助，但是这表明，如果病人出院之后的医学干预是以解决病人的心理问题为重心，可能会更有利于他们的身体健康。这样就可以提供更好的健康服务，降低再入院率和医疗成本。这个相关关系是机器从一大堆数据中筛选出来的，也是人类可能永远都发现不了的。微软不控制数据，这些数据只属于医院；微软没有出彩的想法，那并不是这里需要的东西，相反，微软只是提供了分析工具，也就是Amalga系统来帮助发现有价值的信息。

大数据拥有者依靠技术专家来挖掘数据的价值。但是，虽然受到了高度的赞扬，而且同时拥有“数据武士”这样时髦的名字，但技术专家并没有想象中那么耀眼。他们在大数据中淘金，发现了金银珠宝，可是最后却要把这些财富拱手让给大数据拥有者。

大数据思维公司和个人

第三种类型是有着大数据思维的公司和个人。他们的优势在于，他们能先人一步发现机遇，尽管本身并不拥有数据也不具备专业技能。事实上，很可能正因为他们是外行人，不具备这些特点，他们的思维才能不受限制。他们思考的只有可能，而不考虑所谓的可行。

大数据先锋

FlightCaster的大数据思维

布拉德福德·克罗斯（Bradford　Cross）用拟人手法解释了什么是有大数据思维。2009年8月，也就是在他20多岁的时候，他和四个朋友一起创办了FlightCaster.com。和FlyOnTime.us类似，这个网站致力于预测航班是否会晚点。它主要基于分析过去十年里每个航班的情况，然后将其与过去和现实的天气情况进行匹配。

有趣的是，数据拥有者就做不到这样的事情。因为数据拥有者没有这样使用数据的动机和强制要求。事实上，如果美国运输统计局、美国联邦航空局和美国天气服务这些数据拥有者敢将航班晚点预测用作商业用途的话，国会可能就会举办听证会并否决这个提议。所以使用数据的任务就落到了一群不羁的数学才子的身上。同样，航空公司不可以这么做，也不会这么做，因为这些数据所表达的信息越隐蔽对它们就越有利。FlightCaster的预测是如此的准确，就连航空公司的职员也开始使用它了。但是需要注意的一点就是，虽然航空公司是信息的源头，但是不到最后一秒它是不会公布航班晚点的，所以它的信息是不及时的。

因为有着大数据思维，克罗斯和他的FlightCaster是第一个行动起来的，但也没比别人快多少。所谓大数据思维，是指一种意识，认为公开的数据一旦处理得当就能为千百万人急需解决的问题提供答案。2009年8月，FlightCaster公开发布了。同一个月，FlyOnTime.us的计算机专家们也开始搜刮公开的数据建立他们的网站。最终，FlightCaster的优势慢慢地减弱了。2011年1月，克罗斯和他的同伴把网站卖给了Next Jump，这是一个使用大数据技术进行企业折扣管理的公司。

之后，克罗斯把他的目光转向了另外一个夕阳行业——新闻行业。他发现，这里是一个创新型的外行人可以大有作为的宝地。他的科技创新公司Prismatic收集网上资源并排序，这种排序建立在文本分析、用户喜好、社交网络普及和大数据分析的基础之上。重要的是，这个系统并不介意这是一个青少年的博客、一个企业网站还是《华盛顿邮报》上的一篇报道，只要它的内容相关并且很受欢迎就能排在很靠前的位置。而关于是否受欢迎，是通过它的点击率和分享次数来体现的。

作为一项服务，Prismatic关注的是年青一代与媒体进行交流的新方法，信息的来源并不重要。同时，这也给那些自视过高的主流媒体提了一个醒：公众的力量要远远超过它们，而西装革履的记者们也需要与一群不修边幅的博主进行竞争。也许最令人无法想象的是，Prismatic居然

是从新闻领域内部诞生出来的，虽然它确实收集了大量的数据。美国国家记者俱乐部（National Press Club）的常客从来没有想过要再利用网上的媒体资源，阿蒙克、纽约和印度班加罗尔的分析专家们也没有想过要用这种方法来使用数据。克罗斯顶着一头蓬松的头发，说话吞吞吐吐，可就是这样一个不起眼的外行人，想到了也做到了，他使用这些数据来告诉世界什么是比《纽约时报》更有用的信息来源。

大数据思维这个概念以及一个拥有创新思维的人的地位，与20世纪90年代电子商务初期出现的情况是不一样的。电子商务先驱者们的思想没有被传统行业的固有思维和制度缺陷所限制，因此，在对冲基金工作的金融工程师杰夫·贝索斯创建了网上书店亚马逊而不是巴诺书店[2]；软件开发工程师皮埃尔·奥米迪亚（Pierre Omidyar）开发了一个拍卖网站而不是苏富比（Sotheby's）[3]。如今，拥有大数据思维的领导者通常自己并不拥有数据资源。但就是因为这样，他们不会受既得利益和金钱欲望这样的因素影响而阻碍自己的想法实践。

就像我们看到的，也有公司集合了大数据的多数特点。埃齐奥尼和克罗斯不仅比别人早一步有了这些决胜的思想，他们也有技术优势。Teradata和埃森哲的员工不仅规规矩矩地打卡上班，还时不时会有些机灵的点子。这些原型都有助于我们认识不同公司所承担的角色。我们在上一章节中提到的手机公司掌握了海量的数据却不知道该如何使用，然而，它们可以把这些数据授权给有能力挖掘出数据价值的人。同样地，Twitter一早就决定把它所掌握的海量数据授权给了两家公司。如今的大数据先驱者们通常都有着交叉学科背景，他们会将这些知识与自己所掌握的数据技术相结合，应用于广泛的领域之中。新一代的天使投资人和企业家正在诞生，他们主要是来自谷歌已经离职的员工和所谓的“Paypal黑手党”[4]。他们与少量的计算机科学家一起充当了当今许多数据科技公司的最大靠山。这种将企业和个人置于大数据价值链中的创新性想法促使我们重新审视公司的存在价值。比方说，Salesforce不再是一个单纯为企业提供应用软件的平台，它还能挖掘这些软件所收集到的数据并且释放出它们的巨大价值。

大数据先锋

谷歌与亚马逊，三者兼备

有些比较幸运的公司就有计划地同时涉足了这三个方面。一个很典型的例子就是谷歌，它收集搜索时拼写错误的数据，它也有利用这些数

据创建一个世界上最好的拼写检查程序的好点子，同时它自身也具备挖掘数据价值的技术。谷歌在大数据价值链中同时充当的这三个不同的角色，与谷歌其他项目整合后为谷歌带来了巨大的利润。除此之外，谷歌还通过应用程序接口（APIs）把它掌握的部分数据授权别人使用，这样数据就能重复使用还可以产生附加价值。谷歌地图就是这样，它免费给互联网上的任何人提供服务（尽管访问量很大的网站是需要付费的）。

同样，亚马逊也是数据、技能、思维三者兼备。事实上，该公司的商业模式就是按这个顺序确定的，虽然这与常规不符。刚开始的时候，关于它备受赞誉的推荐系统，亚马逊只有一个初步的想法。它在其1997年的股票市场简介中首先描述了“协同过滤”，这发生在它找到实施这个想法的方法和配备足够的数据资源之前。

虽然谷歌和亚马逊都是三者兼具，但是它们的商业策略并不相同。谷歌在刚开始收集数据的时候，就已经带有多次使用数据的想法。比方说，它的街景采集车收集全球定位系统数据不光是为了创建谷歌地图，也是为了制成全自动汽车。相对地，亚马逊更关注的是数据的基本用途而且也只把数据的二级用途作为额外收益。比方说，它的推荐系统把用户浏览过的网页数据作为线索，但是它并没有利用它预测经济状况和流感爆发。

亚马逊的Kindle电子书阅读器记录了一些读者反复标注和强调过的内容，但是亚马逊并没有把这些数据信息卖给作者或是出版社。书商肯定很乐意知道哪些段落是受读者喜欢的，因为这样他们就能提高销量；作者应该也想知道书籍的哪些地方不受读者欢迎，这样他们就能根据读者的喜好提高作品质量；出版社则可以通过这些数据知道哪些主题的书籍更有可能成为畅销书。但是，亚马逊把这些数据都雪藏了。

一旦得以有效利用，大数据就可以变革公司的赢利模式和传统交流方式。我们举一个典型的例子，通过得到竞争对手所没有的行业信息，欧洲一家汽车制造商重新定位了与它的一个零件供应商的关系。[5]

如今的汽车装满了芯片、传感器和各种软件，一经启动，它们就会及时把汽车状况信息发送到制造商的电脑上。一个典型的中档车大概有60个微型处理器，车上电子仪器的价值占了车辆总价值的三分之一。车载电子仪器之多使汽车成了“漂浮的观景台”，这本是莫里用来形容船舶的。而这些设备监控到的汽车零部件的工作状况，能够在整合之后用来提高汽车的质量，因此，能够掌握这些数据的公司拥有非常大的竞争优

势。

汽车制造商通过与行业外的数据分析公司合作发现，德国供货商供应的油箱的蒸汽泄漏检测传感器存在一些问题，它会对好的油箱产生错误报警达16次。汽车制造商可以把这些信息反馈给供货商要求修理。在商业环境更加和谐的情况下，也许会发生上面说到的情况，但是既然汽车制造商已经在这个项目上花费了一大笔钱，它就会利用这个数据挽回一点点损失。

所以，汽车制造商开始考虑到底应该怎么做：卖掉这个数据？它值多少钱呢？如果供货商推卸责任呢？如果是我自己在操作过程中出现了失误呢？而且它知道，一旦公布了信息，和自己用同样零件的竞争对手也会改进他们的车。更明智的选择应该是，这些数据只能让自己受益，自己的汽车能够有所改进。最终，汽车制造商想到了一个好主意。它通过改进软件而改进了这个零件，而且为这次改进申请了专利。然后，它把这项专利卖给了供货商，价格是很长一段时间内进行数据分析的成本的总额。

全新的数据中间商

谁在这个大数据价值链中获益最大呢？现在看来，应该是那些拥有大数据思维或者说创新性思维的人。就像我们所见的一样，自从信息时代以来，这些第一个吃螃蟹的人都发了大财。但是，这种先决优势并不能维持很长的时间。随着大数据时代的推进，别人也会吸收这种思维，然后那些先驱者的优势就会逐渐减弱。

那么，核心价值会不会在技术上？毕竟，一个金矿的价值也只有在它被挖掘出来之后才有意义。但是，计算机的历史却否定了这一想法。如今，在数据库管理、数据科学、数据分析、机器学习算法等类似行业的技能确实很走俏。但是，随着大数据成为人们生活的一部分，而大数据工具变得更容易和更方便使用，越来越多的人会掌握这些技能，所以这些技能的价值就会相对减少，就像20世纪60~80年代之间计算机编程技术变得越来越普遍一样。现在，国外的外包公司使得基础的计算机编程技术越来越廉价，如今它甚至成为了世界贫困人口的致富驱动力，而不再代表着高端技术。当然，这一切并不是要说大数据技能不重要，只是这不是大数据价值的最主要来源。毕竟，技术是外在的力量。

大数据洞察

现今，我们正处在大数据时代的早期，思维和技能是最有价值的，但是最终，大部分的价值还是必须从数据本身中挖掘。因为在未来，我们可以利用数据做更多的事情，而数据拥有者们也会真正意识到他们所拥有的财富。因此，他们可能会把他们手中所拥有的数据抓得更紧，也会以更高的价格将其出售。继续用金矿来打比方：只有金子才是真正值钱的。

然而，如果数据拥有者做长远打算的话，有一个小问题十分值得关注：那就是在有些情况下会出现“数据中间人”，它们会从各种地方搜集数据进行整合，然后再提取有用的信息进行利用。数据拥有者可以让中间人充当这样的角色，因为有些数据的价值只能通过中间人来挖掘。

大数据先锋

数据中间商，交通数据处理公司Inrix

总部位于西雅图的交通数据处理公司Inrix就是一个很好的例子。它汇集了来自美洲和欧洲近1亿辆汽车的实时交通数据。这些数据来自宝马、福特、丰田等私家车，还有一些商用车，比如出租车和货车。私家车主的移动电话也是数据的来源。这也解释了为什么它要建立一个免费的智能手机应用程序，因为一方面它可以为用户提供免费的交通信息，另一方面它自己就得到了同步的数据。Inrix通过把这些数据与历史交通数据进行比对，再考虑进天气和其他诸如当地时事等信息来预测交通状况。数据软件分析出的结果会被同步到汽车卫星导航系统中，政府部门和商用车队都会使用它。

Inrix是典型的独立运作的大数据中间商。它汇聚了来自很多汽车制造商的数据，这些数据能产生的价值要远远超过它们被单独利用时的价值。每个汽车制造商可能都会利用它们的车辆在行驶过程中产生的成千上万条数据来预测交通状况，这种预测不是很准确也并不全面。但是随着数据量的激增，预测结果会越来越准确。同样，这些汽车制造商并不一定掌握了分析数据的技能，它们的强项是造车，而不是分析泊松分布。所以它们都愿意第三方来做这个预测的事情。另外，虽然交通状况分析对驾驶员来说非常重要，但是这几乎不会影响到一个人是否会购车。所以，这些同行业的竞争者们并不介意通过行业外的中间商汇聚它们手里的数据。

当然，很多行业已经有过信息共享了，比较著名的有保险商实验

室，还有一些已经联网了的行业，比如银行业、能源和通信行业。在这些行业里，信息交流是避免问题最重要的一环，监管部门也要求它们信息互通。市场研究公司把几十年来的数据都汇集在一起，就像一些专门负责审计报刊发行量的公司一样。这是一些行业联盟组织的主要职责。

如今不同的是，数据开始进入市场了。数据不再是单纯意义上的数据，它被挖掘出了新的价值。比方说，Inrix收集的交通状况数据信息会比表面看上去有用得多，它被用来评测一个地方的经济情况，因为它也可以提供关于失业率、零售额、业余活动的信息。2011年，美国经济复苏开始放缓，虽然政客们强烈否定，但是这个信息还是被交通状况分析给披露了出来。Inrix的分析发现，上下班高峰时期的交通状况变好了，这也就说明失业率增加了，经济状况变差了。同时，Inrix把它收集到的数据卖给了一个投资基金，这个投资基金把交通情况视作一个大型零售商场销量的代表，一旦附近车辆很多，就说明商场的销量会增加。在商场的季度财政报表公布之前，这项基金还利用这些数据分析结果换得了商场的一部分股份。

大数据价值链上还出现了很多这样的中间人。比较早期的一个就是Hitwise，现在它已经被益百利收购了。Hitwise与一些互联网服务公司合作，它支付给这些公司一些费用以使用它们的数据。这些数据只是以一个固定的低价授权给Hitwise，而不是按它所得利润的比例抽成。这样一来，Hitwise作为中间人就得到了大部分的利润。另一个中间人的例子就是Quantcast，它通过帮助网站记录用户的网页浏览历史来测评用户的年龄、收入、喜好等个人信息，然后向用户发送有针对性的定向广告。它提供了一个在线系统，网站通过这个系统就能记录用户的浏览情况，而Quantcast就能得到这些数据来帮助自己提高定向广告的效率。

这些中间人在这个价值链中站在了一个收益丰厚的位置上，但是它们并没有威胁到为他们提供数据的数据拥有者的利润。现在，广告业是一个高利润行业，因为大部分的数据都藏身于此，而社会各行各业都急切地需要通过挖掘这些数据进行定向广告。随着越来越多的事情被数据化，越来越多的行业意识到它们与数据有交流，这些独立的数据中间人也会在别处出现。

有时，这些中间人不一定是商业性质的组织，也可能是非营利性的，比如，2011年由美国几个最大的医疗保险公司联合创立的卫生保健成本协会（Health Care Cost Institute）。它们的数据汇集了来自3300万人的50亿份保单，当然这都是匿名的。数据共享之后，这些公司可以看

到在一个较小的独立数据库里看不到的信息。2008年9月，这个超大型数据库就有了第一个重大发现，那就是美国的医疗花费比通货膨胀率的增长速度快3倍之多。但是在各个细微方面的情况就各有不同了：其中急诊室治疗费用上涨了11%，而护理设施的价格实际上是下跌了的。显然，医疗保险公司是不可能把它的价格数据给除非营利性机构之外的任何组织的。这个组织的动机更明确，运行更透明化且更富有责任心。

大数据公司的多样性表明了数据价值的转移。在Decide.com的案例中，产品价格和新产品的发布数据都是由合作的网站提供的，然后合作双方共同分享利润。Decide.com通过人们在这些网站购买产品而赚取佣金，同时提供这些数据的公司也取得了部分利润。相比ITA提供给Farecast的数据不抽取佣金而只是收取基本授权费用的情况，这说明了这个行业的逐渐成熟——如今数据提供者会更占优势。不难想象，埃齐奥尼的下一个科技公司应该就会自己收集数据了，因为数据的价值已经从技术转移到了数据自身和大数据思维上。

随着数据价值转移到数据拥有者手上，传统的商业模式也被颠覆了。上文提到的与供货商进行知识产权交易的欧洲汽车制造商就拥有一个非常专业的数据分析团队，但是还需要一个科技公司来替它挖掘数据的价值。这个科技公司肯定是可以得到报酬的，但是大头还是被这个汽车制造商赚走了。不过，这个科技公司发现了商机，于是它改变了它的商业模式：它为客户承担一定的风险，因为有风险就有回报。而且，它用部分报酬换取了一部分的分析结果，因为这个分析结果是可以循环使用的。比如，对于汽车配件供应商来说，它们未来肯定都想为它们的产品加上测试仪或者把提供产品评估数据写进销售合同的标准条款中，这样它们就能随时改进产品的质量了。

对于中间商来说，公司之间不愿意进行数据共享的问题会让他们感到很头疼。比如Inrix就不再只收集关于地理位置的数据了。2012年，它就关于车辆的自动制动系统何时何地会生效进行了分析，因为有一家汽车制造商用它的遥感勘测系统实时地收集了这些数据。它们认为如果车辆的自动制动系统在某段路上老是启动的话，就说明这段路比较危险，应该考虑更换路径。所以Inrix不仅能够推荐最便捷的路径，而且可以推荐最安全的路径。但是这个制造商并不想和别人分享这些数据，也不愿分享它的全球定位系统收集到的数据。相反，它要求Inrix只能在它生产的车上安装这个系统。在制造商看来，公开这些数据似乎比汇聚众人的数据一起来提高系统的整体精确性更有价值。但即便如此Inrix也相信，

到最后，所有的汽车制造商都会意识到数据共享的好处。Inrix有一种强烈的乐观精神：作为一个数据中间商，它的运行完全是依靠多种多样的数据来源。

大数据时代中的公司正在体验着不同的商业模式。作为中间商的Inrix把它的工作重心放在了设计上，这与众多科技创业公司的商业模式不同。微软掌握着技术的核心专利，但是它却认为，一个独立的小公司可能更容易被接受，更有利于汇聚行业内各方的数据并从知识产权中获利最大。还有，微软用来分析病患再入住率的Amalga系统曾经就是华盛顿中心医院自己的内部急症室软件Azyxxi，这是医院在2006年卖给微软公司的，因为考虑到微软更有能力把这个软件做好和挖掘出这些数据的潜在价值。

2010年UPS就把它的UPS Logistics Technologies部门卖给了一家叫Thoma Bravo的私人股本公司。如今，它已经变成了Roadnet Technologies，可以为多家公司进行线路分析。Roadnet从客户手中收集大量数据，同时为UPS和它的竞争者提供行业内广受认可的标杆性服务。Roadnet的首席执行官兰·肯尼迪（Len Kennedy）解释说，“如果是UPS Logistics，那么UPS的竞争对手肯定不会交出它们的数据，因此，只有让它变成一个独立的公司，UPS的竞争对手才会愿意拿出它们的数据。”最终，每个公司都从中受益了，因为数据汇集之后，系统的精确性就更高了。

认为数据自身而不是技术和思维更值钱的想法，在大数据时代的多笔商业交易中都有所体现。2006年，微软以1.1亿美元的价格购买了埃齐奥尼的大数据公司Farecast。而两年后，谷歌以7亿美元的价格购买了为Farecast提供数据的ITA Software公司。

专家的消亡与数据科学家的崛起

在《点球成金》这部关于奥克兰运动家棒球队如何通过利用统计学和数学建模的方式分析数字，从而取得最终胜利的电影中，有一个有趣的场景，就是灰头发的老球探们坐在一旁评论球员。观众不得不因此感到畏缩，不仅因为它体现了人类做决定时完全不依靠数据的草率，而且因为我们都经历过这种依赖情感而不是科学进行判断的情况。

一个球探说，“他不错，有天赋……而且长得也不错。”

一个满头白发、戴着助听器的老人虚弱地附和道，“他击打动作不错，球一被碰到就一下子弹出去老远。”

另一个球探也附和说，“击打很大声。”

有一个球探打断了对话，说，“他女朋友真丑。”

会议的负责人说，“那是什么意思？”

那个人似乎很肯定地说，“女朋友丑说明没自信呀！”

“很好！”负责人对回答很满意，然后会议继续。

开了一会玩笑之后，一个一直没说话的球探说，“这个人有很大的气场。我的意思是，他还没上场呢，对手就已经提前感受到了他的气势。”

另一个人附和道，“他通过了长相测试，长得不错。他随时都能打球，只是需要点儿上场时间。”

那个常年持不赞同意见的人反复说，“我就是说说，他的女朋友真是长相平平。”

这个场景完全展示了人类判断的误区。一个似乎经过了理智讨论的事情其实是在没有什么实际标准的情况下做出的决定。签约一个几百万美元年薪的球员，也只是看感觉，没有什么客观标准的。是的，这只是电影中的场景，但是生活中这种情况也多得是。这个场景之所以具有讽刺意味，就是因为这是普遍存在的，从曼哈顿的会议室、美国总统办公室到街角咖啡馆，任何地方，这种空泛的推理都到处盛行。

影片《点球成金》改编自迈克尔·刘易斯的《魔球——逆境中制胜的智慧》。讲述的是一个真实的故事，介绍奥克兰运动家棒球队（又称绿帽队或白象队）总经理比利·比恩（Billy Beane）的经营哲学，描写了他抛弃几百年一直依赖的选择球员的传统惯例，采用了一种依靠电脑程序和数学模型分析比赛数据来选择球员的方法。他并没有采用那些像“棒球击球率”这样传统的标准，而是采用了看上去很奇怪的、类似“上垒率”这样的标准。这个方法发现了这项体育赛事的另一面，始终存在却一直被忽略了的一面。一个球员怎样上垒并不要紧，不管是地滚

球还是三垒跑，只要他上垒了就够了。当数据表明偷垒不实用的时候，即使这会让比赛更有看头，比利·比恩也不会再关注这种华而不实的技能。

在一片批评与质疑声中，比恩的“赛伯计量学”（Sabermetrics）在奥克兰运动家棒球队的办公室里被铭记了下来，这是以体育新闻记者比尔·詹姆斯（Bill James）在美国高级棒球研究协会（*Society for Advanced Baseball Research*）中的工作命名的。直到现在，美国高级棒球研究协会一直是一种奇特的亚文化的中心。比恩打破一切常规惯例，就如同伽利略用“太阳中心论”来挑战天主教的权威一样。最终，比恩带领这支备受争议的球队在2002年的美国联盟西部赛中夺得冠军，还取得了20场连胜的战绩。从那以后，统计学家取代球探成为了棒球专家，很多其他球队也开始争相采用“赛伯计量学”来指导球队运作。

同样地，人类从依靠自身判断做决定到依靠数据做决定的转变，也是大数据做出的最大贡献之一。行业专家和技术专家的光芒都会因为统计学家和数据分析家的出现而变暗，因为后者不受旧观念的影响，能够聆听数据发出的声音。他们的判断建立在相关关系的基础上，没有受到偏见和成见的影响，这就如同莫里中校不把干瘦的船长在酒吧喝酒时所说的航道信息当真一样。他们的判断完全依赖于汇集起来的数据所显示出的实际信息，所以有着牢靠的根基。莫里所采用的方法并没有解释风向和水流为什么是这样的原因，但是对于想安全航海的航海家来说，“什么”和“哪里”比“为什么”更加重要。

如今，我们正在见证专家在各个领域影响力的减弱。在传媒界，如“赫芬顿邮报”（*Huffington Post*）和高客网（Gawker）这些网站上传播的新闻通常取决于数据，而不再取决于编辑的新闻敏感度。数据比有经验的记者更能揭示出哪些是符合大众口味的新闻。Coursera，一家网上教育公司，深度地研究它收集的所有数据，比如学生重放过讲座视频的哪个片段，从而找出不明确或者很吸引人的地方，然后反馈给设计课程的团队。这在以前是做不到的，所以老师的教育方法一定会改变。就像我们在前文提到过的，当贝索斯发现算法推荐能促进销量增加的时候，他就不再使用公司的书籍评论员了。

这都意味着，与时俱进才是在职业领域取得成功的必备技能；这样的员工能随时满足公司对他们的期望。安大略的麦格雷戈医生不需要是医院里最好的医生，也不需要是产前护理的世界权威，就能给早产儿提供极好的治疗，因为她采用的治疗方法是电脑在处理了近十年的病患记

录数据之后推荐的。事实上，她也有计算机科学专业的博士学位。

正如我们所见，大数据的先锋们通常并不来自于他们做出了极大贡献的领域。他们是数据分析家、人工智能专家、数学家或者统计学家，但是他们把他们所掌握的技能运用到了各个领域。Kaggle的首席执行官安东尼·戈德布鲁姆（Anthony Goldbloom）说，在这个大数据项目竞赛平台上取得胜利的人通常不来自于他们做出成绩的领域。

一个英国物理学家设计了一个算法系统来预测保险索赔和发现二手车的质量问题，这个系统差点就获胜了；还有一个新加坡的精算师在一个预测人体对化合物的生理反应项目中取得了胜利；同时，在谷歌的机器翻译团队中，这些工程师们都不会说他们翻译出的语言；类似的还有，微软机器翻译部门的统计学家们在茶余饭后的谈资就是说每次一有语言学家离开他们团队，翻译的质量就会变好一点。

当然，行业专家是不会真正消亡的，只是他们的主导地位会发生改变。未来，大数据人才会与他们一样身居高位，就像趾高气扬的因果关系必须与卑微的相关关系分享它的光芒一样。这改变了我们怎样看待知识的价值，因为我们往往倾向于把专业人才看得比全才更重要，也就是说深度就是财富。然而，专业技能就像精确性一样，只适用于“小数据”时代，当时人类掌握的数据永远不够多也不够准确，所以需要依赖直觉和经验指导。在那个时代，经验是先决的，因为只有通过这种无法从书本上和别人口中得到的、埋藏在潜意识里的知识的积累，我们才能做出更明智的决定。

但是当你遭遇海量数据的时候，你就能通过挖掘数据而得到更多。所以大数据分析家会把过去看成是迷信和成规，这不是因为他们更聪明，而是因为他们拥有了这个财富之源——数据。同时，作为外行人，他们不会被行业内的争论所限制，因为他们不会被自己所支持一方的观点所影响而产生偏见，这是他们与行业专家不一样的地方。这一切都意味着，一个员工是否对公司有贡献的判断标准改变了。这也就意味着，你要学的东西、你要了解的人，你要为你的职业生涯所做的准备都改变了。

数学和统计学知识，甚至是有少许编程和网络科学的知识将会成为现代工厂的基础，一如百年前的计算能力或者更早之前的文学。人类的价值将不再体现在与思维类似的同行的交际上，而体现在与各行各业的人的交际上，因为这样知识就能广泛而深刻地进行传播。过去，要成为

一个优秀的生物学家就需要认识很多生物学家，这并没有完全改变。但是如今，不只是专业技能的深度很重要，大数据的广度也变得很重要。要想解决一个生物难题，或许与天体物理学家或者数据视图设计师联系就可以实现。

在电子游戏领域，大数据的普通人才早已经和高级专家站在了一起，他们正在一同改变这个行业。该行业每年收入近100亿美元，比好莱坞的票房收入还要多。过去，游戏公司会设计一个游戏，发布它，指望它能一炮而红。然后，公司会考虑到销售情况，要么继续推出升级版，要么开始研发新游戏。游戏的速度、人物、情节、物品和事件的设定都是基于设计师的创造力，这些设计师对待工作的认真程度就像米开朗基罗画西斯廷教堂时一样。但是，这是一门艺术而不是科学，艺术讲究的是直觉和情感，就像《点球成金》中球探们所表现的一样，然而那个时代已经过去了。

zynga的FarmVille，FrontierVille，FishVille和其他网络游戏都是交互式游戏。表面上，这些游戏允许zynga收集用户数据以及在这些数据的基础上对游戏进行修改，而事实上，这些游戏远远不止一个版本。因为公司可以收集到游戏中的数据，所以一旦有玩家难以过关或者因为某一关不对劲而不想再玩了的时候，zynga就能通过这些数据发现问题，然后对游戏进行修改；但是更加隐性的是，该公司会针对不同的玩家设计不同的游戏，像FarmVille就有好几百个版本。

这个公司的大数据分析家们通过颜色或者是否有玩家看到他的朋友正在使用这些产品，来研究虚拟产品的销量是否增加了。比方说，当数据显示FishVille的玩家购买透明鱼的数量是其他产品的6倍的时候，zynga就会通过多出售透明鱼而谋取更高利润。在Mafia Wars中，数据则显示玩家更喜欢购买有金边的武器和纯白的宠物老虎。这些都不是一个游戏设计师在工作室里能发现的东西，但是数据就能把这些信息传递出来。zynga的首席分析师肯·鲁丁说道，“我们打着游戏公司的幌子，实际上在做的是分析公司的事。我们的运作都是以数据为基础的。”

这种转变意义非凡。大部分人往往都通过经验、回忆以及猜测做决定，就像W.H.奥登（Wystan Hugh Auden）的名诗中所说的“知识退化成骚乱的主观臆想，那是太阳神经丛的感情引起的营养不足”。坐落于马萨诸塞州的巴布森学院商科教授托马斯·达文波特（Thomas Davenport）是多部数据分析著作的作者，他把这种情感称为“黄金般的直觉”。执行官们信任自己的直觉，所以由着它做决定。但是，随着管理决策越来越

受预测性分析和大数据分析的影响和控制，依靠直觉做决定的情况将会被彻底改变。

大数据先锋

The-Numbers.com与电影票房预测

比方说，The-Numbers.com在好莱坞电影上映之前，就能利用海量数据和特定算法预测出一部电影的票房，而这些信息就可以为电影制片人所用。该公司拥有一个包括了过去几十年美国所有商业电影大约3000万条记录的数据库；数据库里有所有关于预算、电影流派、拍摄、阵容、获得奖项和收入等数据。电影的收入是指在北美和全球的票房、海外版权销售收入、影碟销售收入以及租金等。公司创始人兼总裁布鲁斯·纳什（Bruce Nash）说，我们公司开发了一个网络系统，其中有100万条类似“A编剧曾与B导演合作过，C导演曾与D演员合作过”这样的联系信息。

该公司通过找出这样复杂的相关关系来预测电影的收入。借助于这个预测，电影制片人可以向工作室或投资人募资。The-Numbers.com甚至可以告诉客户改变哪些选择可以增收或者降低风险。一次，它的分析发现有一部电影要是启用获得过奥斯卡提名的、身价在500万美元左右的某位一线演员做男一号的话，更有可能票房大卖。还有一次，纳什告诉IMAX工作室，一部航海纪录片需要把预算从1200万美元减少至800万才能赢利。纳什开玩笑地说：“这可乐坏了制片人，但是导演就不高兴了。”

从是否出品一部电影到签下哪个三垒手，公司的决策过程已经有了本质且明显的改变。麻省理工学院商学院教授埃里克·布伦乔尔森（Erik Brynjolfsson）和他的同事一起进行了一项研究，发现决策依赖数据的公司的运营情况比不重视数据的公司出色很多——这些公司的生产率比不使用数据进行决策的公司高6%。这是一个重要的竞争力，虽然随着大数据手段被越来越多的公司采用，这种竞争力会慢慢削弱。

大数据，决定企业竞争力

大数据成为许多公司竞争力的来源，从而使整个行业结构都改变了。当然，每个公司的情况各有不同。大公司和小公司最有可能成为赢家，而大部分中等规模的公司则可能无法在这次行业调整中尝到甜头。

虽然像亚马逊和谷歌一样的行业领头羊会一直保持领先地位，但是和工业时代不一样，它们的企业竞争力并不是体现在庞大的生产规模上。已经拥有的技术配备规模固然很重要，但那也不是它们的核心竞争力，毕竟如今已经能够快速而廉价地进行大量的数据存储和处理了。公司可以根据实际需要调整它们的计算机技术力量，这样就把固定投入变成了可变投入，同时也削弱了大公司的技术配备规模的优势。

大数据洞察

规模仍然很重要，但是如今重要的是数据的规模，也就是说要掌握大量的数据而且要有能力轻松地获得更多的数据。所以，随着拥有的数据越来越多，大数据拥有者将大放异彩，因为他们可以把这些数据转化为价值。

大数据向小数据时代的赢家以及那些线下大公司（如沃尔玛、联邦快递、宝洁公司、雀巢公司、波音公司）提出了挑战，后者必须意识到大数据的威力然后有策略地收集和使用数据。同时，科技创业公司和新兴行业中的老牌企业也准备收集大量的数据。

在过去十年里，航空发动机制造商劳斯莱斯通过分析产品使用过程中收集到的数据，实现了商业模式的转型。坐落于英格兰德比郡的劳斯莱斯运营中心一直监控着全球范围内超过3700架飞机的引擎运行情况，为的就是能在故障发生之前发现问题。数据帮助劳斯莱斯把简单的制造转变成了有附加价值的商业行为：劳斯莱斯出售发动机，同时通过按时计费的方式提供有偿监控服务（一旦出现问题，还进一步提供维修和更换服务）。如今，民用航空发动机部门大约70%的年收入都是来自其提供服务所赚得的费用。

大数据先锋

苹果，挖出“潜伏”的数据价值

苹果公司进军移动手机行业就是一个很好的例子。在iPhone推出之前，移动运营商从用户手中收集了大量具有潜在价值的数据，但是没能深入挖掘其价值。相反，苹果公司在与运营商签订的合约中规定运营商要提供给它大部分的有用数据。通过来自多个运营商提供的大量数据，苹果公司所得到的关于用户体验的数据比任何一个运营商都要多。苹果公司的规模效益体现在了数据上，而不是固有资产上。

大数据也为小公司带来了机遇。用埃里克教授的话说就是，聪明而灵活的小公司能享受到非固有资产规模带来的好处。这也就是说，它们可能没有很多的固有资产但是存在感非常强，也可以低成本地传播它们的创新成果。重要的是，因为最好的大数据服务都是以创新思维为基础的，所以它们不一定需要大量的原始资本投入。数据可以授权但是不能被占有，数据分析能在云处理平台上快速而且低成本地进行，而授权费用则应从数据带来的利益中抽取一小部分。

大大小小的公司都能从大数据中获利，这个情况很有可能并不只是适用于使用数据的公司，也适用于掌握数据的公司。大数据拥有者想尽办法想增加它们的数据存储量，因为这样能以极小的成本带来更大的利润。首先，它们已经具备了存储和处理数据的基础。其次，数据库的融合能带来特有的价值。最后，数据使用者如果只需要从一人手中购得数据，那将更加省时省力。不过实际情况要远远复杂得多，可能还会有一群处在另一方的数据拥有者（个人）诞生。因为随着数据价值的显现，很多人会想以数据拥有者的身份大展身手，他们收集的数据往往是和自身相关的，比如他们的购物习惯、观影习惯，也许还有医疗数据等。

这使得消费者拥有了比以前更大的权利。消费者可以自行决定把这些数据中的多少授权给哪些公司。当然，不是每个人都只在乎把他的数据卖个高价，很多人愿意免费提供这些数据来换取更好的服务，比如想得到亚马逊更准确的图书推荐。但是对于很大一部分对数据敏感的消费者来说，营销和出售他们的个人信息就像写博客、发Twitter信息和在维基百科检索一样自然。

然而，这一切的发生不只是消费者意识和喜好的转变所能促成的。现在，无论是消费者授权他们的信息还是公司从个人手中购得信息都还过于昂贵和复杂。这很可能会催生出一些中间商，它们从众多消费者手中购得信息，然后卖给公司。如果成本够低，而消费者又足够信任这样的中间商，那么个人数据市场就很有可能诞生，这样个人就成功地成为了数据拥有者。美国麻省理工学院媒体实验室的个人数据分析专家桑迪·彭特兰与人一起创办的ID[3]公司已经在致力于让这种模式变为现实。

只有当这些数据中间商诞生并开始运营，而数据使用者也开始使用这些数据的时候，消费者才能真正变成数据掌握者。如今，消费者在等待足够的设备和适当的数据中间商的出现，在这之前，他们希望自己披露的信息越少越好。总之，一旦条件成熟，消费者就能从真正意义上成为数据掌握者了。

不过，大数据对中等规模的公司帮助并不大。波士顿咨询集团的资深技术和商业顾问菲利浦·埃文斯（Philip Evans）说，超大型的公司占据了规模优势，而小公司则具有灵活性。在传统行业中，中等规模的公司比大公司更有灵活性，比小公司更有规模。但是在大数据时代，一个公司没必要非要达到某种规模才能支付它的生产设备所需投入。大数据公司发现它们可以是一个灵活的小公司并且会很成功（或者会被大数据巨头并购）。

大数据洞察

大数据让处于行业两端的公司受益良多，而中等规模的公司要么向两端转换，要么破产。传统行业最终都会转变为大数据行业，无论是金融服务业、医药行业还是制造业。当然，大数据不会让所有行业的中等规模的公司消亡，但是肯定会给可以被大数据分析所取代的中等规模公司带来巨大的威胁。

大数据也会撼动国家竞争力。当制造业已经大幅转向发展中国家，而大家都争相发展创新行业的时候，工业化国家因为掌握了数据以及大数据技术，所以仍然在全球竞争中占据优势。不幸的是，这个优势很难持续。就像互联网和计算机技术一样，随着世界上的其他国家和地区都开始采用这些技术，西方世界在大数据技术上的领先地位将慢慢消失。对于发达国家的大公司来说，好消息就是大数据会加剧优胜劣汰。所以一旦一个公司掌握了大数据，它不但有可能超过它的对手，还有可能遥遥领先。

大数据洞察

竞争正如火如荼地进行。就像谷歌的检索系统需要用户数据才能完好运行，德国的汽车零件供应商需要反馈的数据来提高它的零件质量，所有的公司都能通过巧妙地挖掘数据价值而获得利益。数据能够优化生产和服务，甚至能催生新的行业。

不过，就算有这么多好处，我们依然有担忧的理由。因为随着大数据能够越来越精确地预测世界的事情以及我们所处的位置，我们可能还没有准备好接受它对我们的隐私和决策过程带来的影响。我们的认知和制度都还不习惯这样一个数据充裕的时代，因为它们都建立在数据稀缺的基础之上。下一个章节，我们将探讨大数据所带来的不良影响。

[1] 其余三个是Amadeus，Travelport and Sabre。——作者注

[2] 全美最大的实体书店。——译者注

[3] 全球最有影响力的线下拍卖行。——译者注

[4] 指的是Paypal公司的前领导人皮特·泰尔（Peter Thiel）、雷德·霍夫曼（Reid Hoffman）以及马克斯·莱文奇恩（Max Levchin）。——作者注

[5] 因为这是来自一个分析师在处理大量数据的基础上得到的结果，而这个结果并没有公开，所以我们在此不方便披露该公司的名字。——作者注

第三部分 大数据时代的管理变革

07 风险：让数据主宰一切的隐忧

我们时刻都暴露在“第三只眼”之下：亚马逊监视着我们的购物习惯，谷歌监视着我们的网页浏览习惯，而微博似乎什么都知道，不仅窃听到了我们心中的“TA”，还有我们的社交关系网。

无处不在的“第三只眼”

1989年，柏林墙倒塌，之前的近40年间，民主德国国家安全局“Stasi”雇用了十万左右的全职间谍，时刻在街上开车监视着成千上万民众的一举一动。他们拆看信件、偷窥银行账户信息、在民众家中安装窃听器并且窃听电话。他们还会让情人、夫妇、父母和孩子相互监视，导致人与人之间丧失了最基本的信任。结果，详细记录普通人最私密生活信息的文件至少包括了3900万张索引卡片和铺开足有113公里长的文档。民主德国是一个史无前例的受到如此全面监控的国家。

德国统一20年之后，更多的个人信息被采集和存储了下来。我们时刻都暴露在“第三只眼”之下，不管我们是在用信用卡支付、打电话还是使用身份证。2007年，英国的一家报社曾讽刺地报道，在乔治·奥威尔创作《一九八四》的地方，也就是他的伦敦公寓外60米范围内，起码有30多架摄像机在监视着他的一举一动。

互联网出现之前，如艾可飞和益百利这样的专业数据收集公司就采集、记录了全球范围内大约几百万人口的数据，而它们提供的每个人的个人数据就多达好几百份。而互联网的出现使得监视变得更容易、成本更低廉也更有用处。如今，已经不只是政府在暗中监视我们了。亚马逊监视着我们的购物习惯，谷歌监视着我们的网页浏览习惯，Twitter窃听到了我们心中的“TA”，Facebook似乎什么都知道，包括我们的社交关系网。

进行大数据分析的人可以轻松地看到大数据的价值潜力，这极大地刺激着他们进一步采集、存储、循环利用我们个人数据的野心。随着存储成本继续暴跌而分析工具越来越先进，采集和存储数据的数量和规模将爆发式地增长。如果说在互联网时代我们的隐私受到了威胁，那么大

数据时代是否会加深这种威胁呢？这就是大数据的不利影响吗？

答案是肯定的。大数据还会带来更多的威胁，毕竟，大数据的核心思想就是用规模剧增来改变现状。我们也将分析它是如何加深对我们隐私的威胁的，同时还将面对一个新的挑战，即运用大数据预测来判断和惩罚人类的潜在行为。这是对公平公正以及自由意志的一种亵渎，同时也轻视了决策过程中深思熟虑的重要性。

除了对隐私和倾向的不良影响，大数据还有一个弊端。我们冒险把罪犯的定罪权放在了数据手中，借以表达我们对数据和我们的分析结果的崇尚，但是这实际上是一种滥用。应用得当，大数据会是我们合理决策过程中的有力武器；倘若运用不当，它就可能会变成权贵用来镇压民众的工具，轻则伤害顾客和员工的利益，重则损害公民的人身安全。我们所冒的风险比想象中还要大。如果在隐私和预测方面对大数据管理不当，或者出现数据分析错误，会导致的不良后果比定制化的在线广告要严重得多。

20世纪，我们见证了太多由于数据利用不合理所导致的惨剧。比如1943年，美国人口普查局递交了地址数据来帮助美国政府拘留日裔美国人（当时它没有提交街道名字和具体街号的数据，居然幻想着这样能保护隐私）；荷兰著名的综合民事记录数据则被纳粹分子用来搜捕犹太人；纳粹集中营里罪犯的前臂上刺青的五位数号码与IBM的霍瑞斯穿孔卡片上的号码是一致的，这一切都表明是数据处理帮助实现了大规模的屠杀。

我们的隐私被二次利用了

我们倾向于从数字数据的增长和奥威尔写《1984》时所处“监视炼狱”的角度去理解大数据给个人隐私带来的威胁。但是事实上，不是所有的数据都包含了个人信息。其实，不管是传感器从炼油厂采集的数据、来自工厂的机器数据、机场的气象数据，还是沙井盖爆炸数据都不包含个人信息。英国石油公司和纽约爱迪生联合电力公司不需要（也不想要）个人信息，就能分析挖掘出他们所需要的数据价值。事实上，这方面的数据分析并不威胁个人隐私。

当然，目前所采集的大部分数据都包含有个人信息，而且存在着各种各样的诱因，让我们想尽办法去采集更多、存储更久、利用更彻底，甚至有的数据表面上并不是个人数据，但是经由大数据处理之后就可以

追溯到个人了。

比方说，如今在美国和欧洲部署的一些智能电表每6秒钟采集一个实时读数，这样一天所得到的数据比过去传统电表收集到的所有数据还要多。因为每个电子设备通电时都会有自己独特的“负荷特征”，比如热水器不同于电脑，而它们与Led大麻生长灯[1]又不一样，所以能源使用情况就能暴露诸如一个人的日常习惯、医疗条件和非法行为这样的个人信息。

然而，我们要探讨的主要是大数据是否改变了这种威胁的性质，而不是是否加剧了这种威胁。如果仅仅是加剧了这种威胁，那么我们现在采用的保护隐私的法律法规依然是有效的，我们只需要付出加倍的努力来确保有效性就可以。然而，倘若威胁的性质已经改变了，我们就需要寻求新的解决方案。

不幸的是，我们的担忧一语中的。大数据的价值不再单纯来源于它的基本用途，而更多源于它的二次利用。这就颠覆了当下隐私保护法以个人为中心的思想：数据收集者必须告知个人，他们收集了哪些数据、作何用途，也必须在收集工作开始之前征得个人的同意。虽然这不是进行合法数据收集的唯一方式，“告知与许可”已经是世界各地执行隐私政策的共识性基础（虽然实际上很多的隐私声明都没有达到效果，但那是另一回事）。

更重要的是，大数据时代，很多数据在收集的时候并无意用作其他用途，而最终却产生了很多创新性的用途。所以，公司无法告知个人尚未想到的用途，而个人亦无法同意这种尚是未知的用途。但是只要没有得到许可，任何包含个人信息的大数据分析都需要向个人征得同意。因此，如果谷歌要使用检索词预测流感的话，必须征得数亿用户的同意，这简直无法想象。就算没有技术障碍，又有哪个公司能负担得起这样的人力物力支出呢？

同样，一开始的时候就要用户同意所有可能的用途，也是不可行的。因为这样一来，“告知与许可”就完全没有意义了。大数据时代，告知与许可这个经过了考验并且可信赖的基石，要么太狭隘，限制了大数据潜在价值的挖掘，要么就太空泛而无法真正地保护个人隐私。

同时，想在大数据时代中用技术方法来保护隐私也是天方夜谭。如果所有人的信息本来都已经在数据库里，那么有意识地避免某些信息

就是此地无银三百两。我们把谷歌街景作为一个例子来看，谷歌的图像采集车在很多国家采集了道路和房屋的图像（以及很多备受争议的数据）。但是，德国媒体和民众强烈地抗议了谷歌的行为，因为民众认为这些图片会帮助黑帮窃贼选择有利可图的目标。有的业主不希望他的房屋或花园出现在这些图片上，顶着巨大的压力，谷歌同意将他们的房屋或花园的影像模糊化。但是这种模糊化却起到了反作用，因为你可以在街景上看到这种有意识的模糊化，对盗贼来说，这又是一个此地无银三百两的例子。

另一条技术途径在大部分情况下也不可行，那就是匿名化。匿名化指的是让所有能揭示个人情况的信息都不出现在数据集里，比方说名字、生日、住址、信用卡号或者社会保险号等。这样一来，这些数据就可以在被分析和共享的同时，不会威胁到任何人的隐私。在小数据时代这样确实可行，但是随着数据量和种类的增多，大数据促进了数据内容的交叉检验。

2006年8月，美国在线（AOL）公布了大量的旧搜索查询数据，本意是希望研究人员能够从中得出有趣的见解。这个数据库是由从3月1日到5月31日之间的65.7万用户的2000万搜索查询记录组成的，整个数据库进行过精心的匿名化——用户名称和地址等个人信息都使用特殊的数字符号进行了代替。这样，研究人员可以把同一个人的所有搜索查询记录联系在一起来分析，而并不包含任何个人信息。

尽管如此，《纽约时报》还是在几天之内通过把“60岁的单身男性”、“有益健康的茶叶”、“利尔本的园丁”等搜索记录综合分析考虑后，发现数据库中的4417749号代表的是佐治亚州利尔本的一个62岁寡妇塞尔玛·阿诺德（Thelma Arnold）。当记者找到她家的时候，这个老人惊叹道：“天呐！我真没想到一直有人在监视我的私人生活。”这引起了公愤，最终美国在线的首席技术官和另外两名员工都被开除了。

事隔仅仅两个月之后，也就是2006年10月，DVD租赁商奈飞公司做了一件差不多的事，就是宣布启动“Netflix Prize”算法竞赛。该公司公布了大约来自50万用户的一亿条租赁记录，并且公开悬赏100万美金，举办一个软件设计大赛来提高他们的电影推荐系统的准确度，胜利的条件是把准确度提高10%。同样，奈飞公司也对数据进行了精心的匿名化处理。然而还是被一个用户认出来了，一个化名“无名氏”的未出柜的同性恋母亲起诉了奈飞公司，她来自保守的美国中西部。

通过把奈飞公司的数据与其他公共数据进行对比分析，得克萨斯大学的研究人员很快发现，匿名用户进行的收视率排名与互联网电影数据库（IMDb）上实名用户所排的是匹配的。

大数据的力量

概括地说，研究发现每对6部不出名的电影进行排序，我们就有84%的概率可以辨认出奈飞公司这个顾客的身份。而如果我们知道这个顾客是哪天进行了排序的话，那么他被从这个50万人的数据库中挑出来的概率就会高达99%。

在美国在线的案例中，我们被我们所搜索的内容出卖了。而奈飞公司的情况则是因为不同来源数据的结合暴露了我们的身份。这两种情况的出现，都是因为公司没有意识到匿名化对大数据的无效性。而出现这种无效性则是由两个因素引起的，一是我们收集到的数据越来越多，二是我们会结合越来越多不同来源的数据。

科罗拉多大学的法学教授保罗·欧姆（Paul Ohm），同时也是研究反匿名化危害的专家，认为针对大数据的反匿名化[2]，现在还没有很好的办法。毕竟，只要有足够的数据，那么无论如何都做不到完全的匿名化。更糟的是，最近的研究表明，不只是传统数据容易受到反匿名化的影响，人们的社交关系图，也就是人们的相互联系也将同受其害。

大数据洞察

在大数据时代，不管是告知与许可、模糊化还是匿名化，这三大隐私保护策略都失效了。如今很多用户都觉得自己的隐私已经受到了威胁，当大数据变得更为普遍的时候，情况将更加不堪设想。

与25年之前的民主德国相比，现在我们所受的监控没有减少，反而变得越来越容易、严密以及低成本。采集个人数据的工具就隐藏在我们日常生活所必备的工具当中，比如网页和智能手机应用程序。我们知道大多数的汽车中都装了一个“黑盒子”——用来监测安全气囊激活的情况，而如今，一旦出现具有争议的交通案件，这个黑盒子所采集的数据就可以在法庭上充当证据。当然，如果企业采集数据只是来提高绩效，我们就不用像被Stasi窃听那样而感到那么害怕。毕竟企业再强大，也不如国家强制力。

不过，即使它们不具备国家强制力，想到各种各样的公司在我们不知情的情况下采集了我们日常生活方方面面的数据，并且进行了数据共享以及一些我们未知的运用，这还是很恐怖的。对大数据大加利用的不只是私营企业，政府也不甘落后。

据《华盛顿邮报》2010年的研究表明，美国国家安全局每天拦截并存储的电子邮件、电话和其他通信记录多达17亿条。前美国安全局官员威廉·宾尼（William Binney）估计政府采集的美国及他国公民的通信互动记录有20万亿次之多，其中包括谁和谁通过话、发过电子邮件、进行过电汇等信息。为了弄明白这所有的数据，美国建立了庞大的数据中心，其中美国国家安全局就耗资12亿美元在犹他州的威廉姆斯堡建立了一个。

如今，不再只是负责反恐的秘密机关需要采集更多的数据，所有的政府部门都需要，所以，数据采集扩展到了金融交易、医疗记录和Facebook状态更新等各个领域，数据量之巨可想而知。政府其实处理不了这么多数据，那为什么要费力采集呢？

这是因为在大数据时代，监控的方式已经改变了。过去，调查员为了尽可能多地知道嫌疑人的信息，需要把鳄鱼夹夹到电话线上。当时最重要的是能深入调查某个人，而现在情况不一样了，比如谷歌和Facebook的理念则是人就是社会关系、网上互动和内容搜索的加和。所以，为了全面调查一个人，调查员需要得到关于这个人的最广泛的信息，不仅是他们认识的人，还包括这些人又认识哪些人等。过去的技术条件没法做到这样的分析，但是今非昔比了。

不过，虽然企业和政府拥有的这种采集个人信息的能力，让我们感到很困扰，但也还是没有大数据所引起的另一个新问题让我们更恐慌，那就是用预测来判断我们。

预测与惩罚，不是因为“所做”，而是因为“将做”

约翰·安德顿（John Anderton）是华盛顿特区警局预防犯罪组的负责人。这是特别的一天，早上，他冲进了住在郊区的霍华德·马克斯（Howard Marks）的家中并逮捕了他，后者打算用剪刀刺杀他的妻子，因为他发现他妻子给他戴了“绿帽子”。安德顿又防止了一起暴力犯罪案件的发生。他大声说：“我以哥伦比亚特区预防犯罪科的名义逮捕你，你即将在今天谋杀你的妻子萨拉·马克斯（Sarah Marks）……”其他的警

察开始控制霍华德，霍华德大喊冤枉，“我什么都没有做啊！”

这是电影《少数派报告》（*Minority Report*）开始时的场景，这部电影描述的是一个未来可以准确预知的世界，而罪犯在实施犯罪前就已受到了惩罚。人们不是因为所做而受到惩罚，而是因为将做，即使他们事实上并没有犯罪。虽然电影中预测依靠的不是数据分析，而是三个超自然人的想象，但是《少数派报告》所描述的这个令人不安的社会正是不受限制的大数据分析可能会导致的：罪责的判定是基于对个人未来行为的预测。

我们已经看到了这种社会模式的萌芽。30多个州的假释委员正使用数据分析来决定是释放还是继续监禁某人。越来越多的美国城市，从洛杉矶的部分地区到整个里士满（美国弗吉尼亚州首府），都采用了“预测警务”（也就是大数据分析）来决定哪些街道、群体还是个人需要更严密的监控，仅仅因为算法系统指出他们更有可能犯罪。

在孟菲斯市，一个名为“蓝色粉碎”[3]的项目为警员提供情报，关于哪些地方更容易发生犯罪事件，什么时候更容易逮到罪犯。这个系统帮助执法部门更好地分配其有限的资源。这个项目自2006年启动以来，孟菲斯的重大财产和暴力犯罪发生率约下降了26%（虽然这与这个项目不一定有因果关系）。

在里士满市的另一个项目中，警察把犯罪数据与其他数据相关联，比方说市里的大公司何时给员工发工资，当地举办音乐会或者运动赛事的时间。这证实了警方对犯罪趋势的预测，有时也会帮助警方推算出更准确的犯罪趋势。例如，里士满市的警察一直觉得在枪击事件之后会出现一个犯罪高峰期，大数据证明了这种想法，但是也发现了一个漏洞，即高峰不是紧随枪击事件而来的，而是两个星期之后才会出现。

这些系统通过预测来预防犯罪，最终要精准到谁会犯罪这个级别。这是大数据的新用途。众多科幻小说的丰富演绎进一步揭示了机场日常安检的平庸和困境。美国国土安全部正在研发一套名为未来行为检测科技（Future Attribute Screening Technology，简称FAST）的安全系统，通过监控个人的生命体征、肢体语言和其他生理模式，发现潜在的恐怖分子。研究者认为，通过监控人类的行为可以发现他们的不良意图。美国国土安全部声称，在研究测试中，系统检测的准确度可以达到70%。（测试方法并不可知，难道是要志愿者假扮恐怖分子，然后看看系统是否能发现他们的不良意图吗？）尽管这些研究还处于早期阶段，执法者

和监管部门还是对其给予了高度重视。

我们可以用大数据来预防犯罪，听起来真不错。毕竟在犯罪发生之前及时制止比事后再惩罚要好得多，不是吗？因为我们避免了犯罪的发生，也就挽救了可能被伤害的人，同时社会整体也受益了。

但是这很危险，因为如果我们可以用大数据来预防犯罪，我们就可能会想进一步惩罚这个未来的罪犯。这也是符合逻辑的，因为我们会觉得如果只是阻止了他的犯罪行为而不采取惩罚措施的话，他就可能因为不受损失而再次犯罪；如果我们因为他未实施的犯罪行为而惩罚他的话，可能就会威慑到他。

基于预测基础上的惩罚似乎也是我们现在惯行方法的一种提升。现代社会是建立在预防不健康、危险和非法行为基础上的。我们为了预防肺癌而减少吸烟率、为了避免在车祸中死亡而系安全带、为了避免被劫机而不允许带枪支登机，所有这些预防措施都限制了我们的自由，但是我们愿意为了防止更大的灾难而做出适当的牺牲。

大多数情况下，我们已经在以预测之名采用大数据分析。它把我们放在一个特定的人群之中来对我们进行界定。保险精算表上指出，超过50岁的男性更容易患前列腺癌，所以你如果不幸正好处于这个年龄段，就需要支付更多的保险费用，即使你根本就没得过这个病。没有高中文凭的人更容易偿还不起债务，所以如果你没有高中文凭，就可能贷不到款或者必须支付更高的保险费。有的人在过安检的时候，可能会需要进行额外的检查，仅仅是因为他带有某种特定的特征。

这都是如今的小数据时代所采用的“画像”背后的指导思想。在一个数据库中找到普遍联系，然后对适用于这种普遍联系的个人深入勘察。这适用于团体内的每个人，是一条普遍规则。当然，“画像”意义颇多，不只意味着对一个特定群体的区分，而且指“牵连犯罪”，不过这是一种滥用，所以“画像”有严重的缺陷。

大数据洞察

大数据替我们规避了“画像”的缺陷，因为大数据区分的是个人而不是群体，所以我们不会再通过“牵连犯罪”给群体中的每个人都定罪。如今，一个用现金购买头等舱单程票的阿拉伯人不会再被认为是恐怖分子而接受额外的检查，只要他身上的其他数据表明他基本没有恐怖主义倾

向。因此，大数据通过给予我们关于个人自身更详尽的数据信息，帮我们规避了“画像”的缺陷——直接将群体特征强加于个人。

其实，我们一直在用“画像”来帮助我们确定个人的罪责，大数据所做的并没有本质的差别，只是让这种方法更完善、更精准、更具体和更个性化。因此，如果大数据预测只是帮助我们预防不良行为，我们似乎是可以接受的。但是，倘若我们使用大数据预测来判定某人有罪并对其尚未实施的行为进行惩罚，就可能让我们陷入一个危险的境地。

基于未来可能行为之上的惩罚是对公平正义的亵渎，因为公平正义的基础是人只有做了某事才需要对它负责。毕竟，想做而未做不是犯罪。社会关于个人责任的基本信条是，人为其选择的行为承担责任。如果有人在被别人用枪威胁的情况下打开了公司的保险柜，他并不需要承担责任，因为他别无选择。

如果大数据分析完全准确，那么我们的未来会被精准地预测，因此在未来，我们不仅会失去选择的权利，而且会按照预测去行动。如果精准的预测成为现实的话，我们也就失去了自由意志，失去了自由选择生活的权利。既然我们别无选择，那么我们也就不需要承担责任，这不是很讽刺吗？

当然，精准的预测是不现实的。大数据分析只能预测一个人未来很有可能进行的行为。

比方说，宾夕法尼亚大学教授理查德·伯克（Richard Berk）建立了一个大数据模型，他声称这个模型可以预测一个判缓刑或者假释的人一旦提前释放会不会再次杀人。他输入了海量的特定案件变量，包括监禁的原因、首次犯罪的时间、年龄、性别等个人数据。伯克说他的模型对未来行为预测的准确率可以达到75%。这听起来似乎还不错。但是，这也意味着如果假释委员会依靠他的分析，就会在每4个人中出现一次失误。

但是，主要的问题并不是出在社会需要面对更多威胁上，而是我们在人们真正犯罪之前对他进行惩罚否定了人的自由权利。我们永远不会知道这个受惩罚的人是否会真正犯罪，因为我们已经通过预测预先制止了这种行为，如此一来，我们就没有让他按照他的意愿去做，但是我们却依然坚持他应该为自己尚未实施的未来行为付出代价，而我们的预测也永远无法得到证实。

这否定了法律系统或者说我们的公平意识的基石——无罪推定原则。因为我们被追究责任，居然是为了我们可能永远都不会实施的行为。对预测到的未来行为判罪也否认了我们进行道德选择的能力。

大数据预测给我们带来的威胁，不仅仅局限于司法公正上，它还会威胁到任何运用大数据预测对我们的未来行为进行罪责判定的领域，比如民事法庭案件中判定过失以及公司解雇员工的决策。

大数据洞察

也许，大数据预测可以为我们打造一个更安全、更高效的社会，但是却否定了我们之所以为人的重要组成部分——自由选择的能力和行为责任自负。大数据成为了集体选择的工具，但也放弃了我们的自由意志。

大数据的不利影响并不是大数据本身的缺陷，而是我们滥用大数据预测所导致的结果。大数据预测是建立在相关性基础上的。让人们为还未实施的未来行为买单是带来不利影响的主要原因，因为我们把个人罪责判定建立在大数据预测的基础上是不合理的。

大数据有利于我们理解现在和预见未来的风险，如此一来，我们就可以相对应地采取应对措施。大数据预测可以帮助患者、保险公司、银行和顾客，但是大数据不能告诉我们因果关系。相对地，进行个人罪责推定需要行为人选择某种特定的行为，他的选择是造成这个行为的原因。但大数据并不是建立在因果关系基础上的，所以它完全不应该用来帮助我们进行个人罪责推定。

麻烦的是，人们习惯性地从因果关系的视角来理解世界。因此，大数据总是被滥用于因果分析，而且我们往往非常乐观地认为，只要有了大数据预测的帮助，我们进行个人罪责判定就会更高效。

这是一个典型的滑坡，可能直接导致《少数派报告》中的情况成为现实——我们将生活在一个没有独立选择和自由意志的社会，在这里我们的道德指标将被预测系统所取代，个人一直受到集体意志的冲击。简单地说，如果一切都成为现实，大数据就会把我们禁锢在可能性之中。

数据独裁

大数据大大地威胁到了我们的隐私和自由，这都是大数据带来的新威胁。但是与此同时，它也加剧了一个旧威胁：过于依赖数据，而数据远远没有我们所想的那么可靠。要揭示大数据分析的不可靠性，恐怕没有比罗伯特·麦克纳马拉（Robert McNamara）的例子更贴切的了。

麦克纳马拉是一个执迷于数据的人。20世纪60年代早期，在越南局势变得紧张的时候，他被任命为美国国防部长。任何事情，只要可以，他都会执意得到数据。他认为，只有运用严谨的统计数据，决策者才能真正理解复杂的事态并做出正确的决定。他眼中的世界就是一堆桀骜不驯的信息的总和，一旦划定、命名、区分和量化之后，就能被人类驯服并加以利用。麦克纳马拉追求真理，而数据恰好能揭示真理。他所掌握的数据中有一份就是“死亡名单”。

麦克纳马拉对数字的执迷从年轻的时候就开始了，当时他还是哈佛商学院的学生，后来，他以24岁的年纪成为了最年轻的副教授。第二次世界大战期间，他把这种严密的数字意识运用到了工作之中，当时他是五角大楼里被称为“统计控制队”中的一名精英，这个队伍让世界权力的中心人物都开始依靠数据进行决策。在这之前，部队一直很盲目。比方说，它们不知道飞机备用零件的种类、数量和放置位置。1943年制作的综合清单为部队节省了36亿美元。现代战争需要资源的合理分配，他们所做的非常了不起。

战争结束的时候，他们决定通力合作拯救濒临倒闭的福特汽车公司。福特二世（Henry Ford Ⅱ）绝望地交出了自己的控制权。就像他们投入战争的时候完全不懂军事一样，这一次，他们也不关心如何制作汽车。但是奇妙的是，这群精明小子居然救活了福特公司。

麦克纳马拉对数据的执迷迅速升温，开始凡事都考虑数据集。工厂经理迅速地生成麦克纳马拉所要求的数字，不管对错。他规定只有在旧车型的所有零件的存货用完之后才能生产新车型，愤怒的生产线经理们一股脑将剩余的零件全部倒进了附近的河里。当前线员工把数据返回的时候，总部的高管们都满意地点了点头，因为规定执行得很到位。但是工厂里盛行一个笑话，是说河面上可以走人了，因为河里有很多1950年或者1951年生产的车型的零件，在河面上走就是在生锈的零件上走。

麦克纳马拉是典型的20世纪经理人——完全依赖数字而非感情的理智型高管，他可以把他的数控理论运用到任何领域。1960年，他被任命为福特汽车公司的总裁，在位只有几周，他就被肯尼迪总统任命为美国

国防部部长。

随着越南战争升级和美军加派部队，这变成了一场意志之战而非领土之争。美军的策略是逼迫越共走上谈判桌。于是，评判战争进度的方法就是看对方的死亡人数。每天报纸都会公布死亡人数。支持战争的人把这作为战争胜利的标志，反战的人把它作为道德沦丧的证据。死亡人数是代表了一个时代的数据集。

1977年，一架直升机从西贡的美国大使馆屋顶上撤离了最后一批美国公民。两年之后，一位退休的将军道格拉斯·金纳德（Douglas Kinnard）发表了《战争管理者》（*The War Managers*）。这是一个关于将军们对越战看法的里程碑式的调查。它揭露了量化的困境。仅仅只有2%的美国将军们认为用死亡人数衡量战争成果是有意义的，而三分之二的人认为大部分情况下数据都被夸大了。一个将军评论称，“那都是假的，完全没有意义”；另一个说道，“公开撒谎”；还有一个将军则认为是像麦克纳马拉这样的人表现出了对数据的极大热忱，导致很多部门一层一层地将数字扩大化了。

就像福特的员工将零件投入河中一样，下级军官为了达成命令或者升官，会汇报可观的数字给他们的上级，只要那是他们的上级希望听到的数字。[4]麦克纳马拉和他身边的人都依赖并且执迷于数据，他认为只有通过电子表格上有序的行、列、计算和图表才能真正了解战场上发生了什么。他认为掌握了数据，也就进一步接近了真理（上帝）。

大数据洞察

美国军方在越战时对数据的使用、滥用和误用给我们提了一个醒，在由“小数据”时代向大数据时代转变的过程中，我们对信息的一些局限性必须给予高度的重视。数据的质量可能会很差；可能是不客观的；可能存在分析错误或者具有误导性；更糟糕的是，数据可能根本达不到量化它的目的。

我们比想象中更容易受到数据的统治——让数据以良莠参半的方式统治我们。其威胁就是，我们可能会完全受限于我们的分析结果，即使这个结果理应受到质疑。或者说，我们会形成一种对数据的执迷，因而仅仅为了收集数据而收集数据，或者赋予数据根本无权得到的信任。

随着越来越多的事物被数据化，决策者和商人所做的第一件事就是

得到更多的数据。“我们相信上帝，除了上帝，其他任何人都必须用数据说话。”这是现代经理人的信仰，也回响在硅谷的办公室、工厂和市政厅的门廊里。善加利用，这是极好的事情，但是一旦出现不合理利用，后果将不堪设想。

教育似乎在走下坡路？用标准化测试来检验学生的表现和评定对老师或学校的奖惩是不合理的。考试是否能全面展示一个学生的能力？是否能有效检测教学质量？是否能反映出一个有创造力、适应能力强的现代师资队伍所需要的品质？这些都饱受争议，但是，数据不会承认这些问题的存在。

如何防止恐怖主义？创造一层层的禁飞名单、阻止任何与恐怖主义有关的个人搭乘飞机，这真的有用吗？回答是：值得怀疑。想想那件非常出名的事情，马萨诸塞州参议员特德·肯尼迪（Ted Kennedy）不就因为仅仅与该数据库中的一个人名字相同而被诱捕、拘留并且调查了吗？

与数据为伴的人可以用一句话来概括这些问题，“错误的前提导致错误的结论。”有时候，是因为用来分析的数据质量不佳；但在大部分情况下，是因为我们误用了数据分析结果。大数据要么会让这些问题高频出现，要么会加剧这些问题导致的不良后果。

我们在书中举过无数谷歌的例子，我们明白它的一切运作都是基于数据基础之上的。很明显，它大部分的成功都是数据造就的，但是偶尔谷歌也会因为数据栽跟头。

谷歌公司的创始人拉里·佩奇和谢尔盖·布林一直强调要得到每个应聘者申请大学时的SAT成绩以及大学毕业时的平均绩点。他们认为，前者能彰显潜能，后者则展现成就。因此，当40多岁、成绩斐然的经理人在应聘时被问到大学成绩的时候，就完全无法理解这种要求。尽管公司内部研究早就表明，工作表现和这些分数根本没有关系，谷歌依然冥顽不化。

谷歌本应该懂得抵制数据的独裁。考试结果可能一生都不会改变，但是它并不能测试出一个人的知识深度，也展示不出一个人的人文素养，学习技能之外，科学和工程知识才是更适合考量的。谷歌在招聘人才方面如此依赖数据让人很是费解，要知道，它的创始人可是接受过注重学习而非分数的蒙台梭利教育。谷歌就是在重蹈前人覆辙，过去美国的科技巨头们也把个人简历看得比个人能力重要。如果按谷歌的做法，

其创始人都没有资格成为传奇性的贝尔实验室的经理，因为他们都在博士阶段辍学了；比尔·盖茨和马克·扎克伯格也都会被淘汰，因为他们都没有大学文凭。

谷歌对数据的依赖有时太夸张了。玛丽莎·迈尔（Marissa Mayer）曾任谷歌高管职位，居然要求员工测试41种蓝色的阴影效果中，哪种被人们使用最频繁，从而决定网页工具栏的颜色。谷歌的数据独裁就是这样达到了顶峰，同时也激起了反抗。

2009年，谷歌首席设计师道格·鲍曼（Doug Bowman）因为受不了随时随地的量化，愤然离职。“最近，我们竟然争辩边框是用3、4还是5倍像素，我居然被要求证明我的选择的正确性。天呐！我没办法在这样的环境中工作，”她离职后在博客上面大发牢骚，“谷歌完全是工程师的天下，所以只会用工程师的观点解决问题——把所有决策简化成一个逻辑问题。数据成为了一切决策的主宰，束缚住了整个公司。”

其实，卓越的才华并不依赖于数据。史蒂夫·乔布斯多年来持续不断地改善Mac笔记本，依赖的可能是行业分析，但是他发行的iPod、iPhone和iPad靠的就不是数据，而是直觉——他依赖于他的第六感。当记者问及乔布斯苹果推出iPad之前做了多少市场调研时，他那个著名的回答是这样的：“没做！消费者没义务去了解自己想要什么。”

詹姆斯·斯科特（James Scott）教授是耶鲁大学政治学和人类学教授，他在《国家的视角》（*Seeing Like a State*）一书中记录了政府如何因为它们对量化和数据的盲目崇尚而陷人民的生活于水深火热之中。

它们使用地图来确定社区重建，却完全不知道其中民众的生活状态。它们使用大量的农收数据来决定采取集体农庄的方式，但是它们完全不懂农业。它们把所有人们一直以来用之交流的不健全和系统的方式按照自己的需求进行改造，只是为了满足可量化规则的需要。在斯科特看来，大数据使用成了权力的武器。

这是数据独裁放大了的写照。同样，也是这种自大导致美国基于死亡人数而不是更理智的衡量标准来扩大越南战争的规模。1976年，在与日俱增的国内压力下，麦克纳马拉在一次演讲中说道，“事实上，真的不是每一个复杂的人类情况都能简化为曲线图上的线条、图表上的百分点或者资产负债表上的数字。但是如果不对可量化的事物进行量化，我们就会失去全面了解该事物的机会。”只要得到了合理的利用，而不单

纯只是为了“数据”而“数据”，大数据就会变成强大的武器。

20世纪70年代，罗伯特·麦克纳马拉一直担任世界银行行长。20世纪80年代，他俨然变成了和平的象征。他为反核武器和环境保护摇旗呐喊。然后，他经历了一次思想的转变并且出版了一本回忆录《回顾：越战的悲剧与教训》（*In Retrospect*：*The Tragedy and Lessons of Vietnam*），书中批判了战争的错误指导思想并承认了他当年的行为“非常错误”，他写道，“我们错了，大错特错！”但书中还是只承认了战争的整体策略的错误，并未具体流露出对数据和“死亡人数”饱含感情的忏悔。他承认统计数据具有“误导或者迷惑性”，“但是对于你能计算的事情，你应该计算；死亡数就属于应该计算的……”2009年，享年93岁的麦克纳马拉去世，他被认为是一个聪明却并不睿智的人。

大数据诱使我们犯下罗伯特·麦克纳马拉所犯的罪行，也让我们盲目信任数据的力量和潜能而忽略了它的局限性。把大数据等同于死亡人数，我们只需要想想上文提到的谷歌流感趋势。设想一下致命的流感正肆虐全国，而这并不是完全不可能出现的；医学专家们会非常感激通过检索词条，我们能够实时预测流感重灾地，他们也就能及时去到最需要他们的地方。

但是在危急时刻，政府领导可能会认为只知道哪里流感疫情最严重还远远不够。如果试图抑制流感的传播，就需要更多的数据。所以他们呼吁大规模的隔离，当然不是说隔离这个地区的所有人，这样既无必要也太费事。大数据能给我们更精确的信息，所以我们只需隔离搜索了和流感有最直接关系的人。如此，我们有了需要隔离的人的数据，联邦特工只需通过IP地址和移动GPS提供的数据，找出该用户并送入隔离中心。

我们可能觉得，这种做法很合理，但是事实上，这是完全错误的。相关性并不意味着有因果关系。通过这种方式找出的人，可能根本就没有感染流感。他们只是被预测所害，更重要的是，他们成了夸大数据作用同时又没有领会数据真谛的人的替罪羊。谷歌流感趋势的核心思想是这些检索词条和流感爆发相关，但是这也可能只是医疗护工在办公室听到有人打喷嚏，然后上网查询如何防止自身感染，而不是因为他们自己真的生病了。

挣脱大数据的困境

大数据为监测我们的生活提供了便利，同时也让保护隐私的法律手段失去了应有的效力。面对大数据，保护隐私的核心技术不再适用了。同样，通过大数据预测，对我们的未来想法而非实际行为采取惩罚措施，也让我们惶恐不安，因为这否认了自由意志并伤害了人类尊严。

同时，那些尝到大数据益处的人，可能会把大数据运用到它不适用的领域，而且可能会过分膨胀对大数据分析结果的信赖。随着大数据预测的改进，我们会越来越想从大数据中掘金，最终导致一种盲目崇拜，毕竟它是如此的无所不能。这就是我们必须从麦克纳马拉的故事中引以为戒的。

必须杜绝对数据的过分依赖，以防我们重蹈伊卡洛斯[5]的覆辙。他就是因为过分相信自己的飞行技术，最终误用了数据而落入了海中。下一章，我们将探讨如何让数据为我们所用，而不让我们成为数据的奴隶。

---

[1] 这是一种植物补光灯，也是植物生长灯的一种，依照植物生长需要太阳光的规律，代替太阳光给植物提供更好的生长发育环境。——编者注

[2] 是指通过技术手段从匿名化数据中挖出用户的真实身份。——译者注

[3] Blue CRUSH，来自Crime Reduction Utilizing Statistical History的缩写，意为利用统计历史减少犯罪。——作者注

[4] 更典型而悲痛的例子是“大跃进”时期各地疯狂虚报粮食产量，以至于中央完全没有意识到20世纪60年代初可怕的饥荒，还在大量出口粮食以换取其他战略物资。——译者注

[5] 希腊神话人物，是希腊神话中代达罗斯的儿子。与代达罗斯使用蜡和羽毛造的双翼逃离克里特岛时，因过于自信，飞得太高，双翼上的蜡遭太阳炙烤融化而跌落水中丧生，被埋葬在一个海岛上。为了纪念伊卡洛斯，埋葬伊卡洛斯的海岛被命名为伊卡利亚。——编者注

08 掌控：责任与自由并举的信息管理

当世界开始迈向大数据时代时，社会也将经历类似的地壳运动。在改变人类基本的生活与思考方式的同时，大数据早已在推动人类信息管理准则的重新定位。然而，不同于印刷革命，我们没有几个世纪的时间去适应，我们也许只有几年时间。

一场管理规范的变革

我们在生产和信息交流方式上的变革必然会引发自我管理所用规范的变革。同时，这些变革也会带动社会需要维护的核心价值观的转变。我们以印刷机的发明导致的信息洪流为例。

1450年前后，古登堡发明了活字印刷机，在这之前，思想的传播受到了极大的限制。一方面，书籍大多被封禁在修道院的图书馆里，依照天主教精心制定的规定，被僧侣严格看守着，为的是确保并维护其统治地位。在教堂之外，少数几所大学也收藏了一些书籍，大概几百本的样子；15世纪初，剑桥大学图书馆大概有122本大部头。另一方面，读写水平的欠缺也是当时信息传播受限的一个重要因素。

古登堡的印刷机让书籍和手册的大量刊印成为可能。马丁·路德（Martin Luther）把拉丁语版本的《圣经》翻译成日常使用的德文，让越来越多的人可以不通过牧师而直接聆听上帝的声音，德语版的《圣经》是当时卖得最好的书，这也让他更确信《圣经》可以印刷、分发给成千上万的人。就这样，信息传播越来越广泛。

这种巨变也使得创立新规范来管理活字印刷术所引发的信息爆炸的条件变得成熟。审查和许可条例被创立，用来规范和管理出版物。著作权法的制定为创作者带来了进行创作的法律和经济动力。随后，保护公民言论自由被写入了宪法。一如既往，权利伴随着责任产生了。当低俗的报纸践踏人们隐私权或诽谤其名誉时，法律规范就会出现以保护人们的隐私权并允许他们对文字诽谤提出上诉。

可是，变革并不止于规范。这种管理规范上的改变也体现了当时更深层次的价值观转变。在古登堡时期，人类第一次意识到了文字的力量；最终，也意识到了信息广泛传播的重要性。几个世纪过去了，我们选择获取更多的信息而非更少，并且借助限制信息滥用的规范而不是最

初的审查来防止其泛滥。

随着世界开始迈向大数据时代，社会也将经历类似的地壳运动。在改变我们许多基本的生活和思考方式的同时，大数据早已在推动我们去重新考虑最基本的准则，包括怎样鼓励其增长以及怎样遏制其潜在威胁。然而，不同于印刷革命，我们没有几个世纪的时间去慢慢适应，我们也许只有几年时间。

大数据时代，对原有规范的修修补补已经满足不了需要，也不足以抑制大数据带来的风险——我们需要全新的制度规范，而不是修改原有规范的适用范围。想要保护个人隐私就需要个人数据处理器对其政策和行为承担更多的责任。同时，我们必须重新定义公正的概念，以确保人类的行为自由（也相应地为这些行为承担责任）。新机构和专家们需要设计复杂的程序对大数据进行解读，挖掘出其潜在的价值和结论。他们也要向那些可能受害于大数据结论的人——因之被剥夺了工作、接受医疗或贷款权利的人，提供支持。对已有的规范进行修修补补已经不够了，我们需要推陈出新。

管理变革1：个人隐私保护，从个人许可到让数据使用者承担责任

数十年来，全球范围内的隐私规范都开始让人们自主决定是否、如何以及经由谁来处理他们的信息，把这种控制权放在了人们自己手中，这也是隐私规范的核心准则。在互联网时代，这个伟大的理想往往会演变成“告知与许可”的公式化系统。

在大数据时代，因为数据的价值很大一部分体现在二级用途上，而收集数据时并未作这种考虑，所以“告知与许可”就不能再起到好的作用了。

在大数据时代，我们需要设立一个不一样的隐私保护模式，这个模式应该更着重于数据使用者为其行为承担责任，而不是将重心放在收集数据之初取得个人同意上。这样一来，使用数据的公司就需要基于其将对个人所造成的影响，对涉及个人数据再利用的行为进行正规评测。当然，并不是说任何时候都必须非常详尽。

未来的隐私保护法应当区分用途，包括不需要或者只需要适当标准化保护的用途。对于一些危险性较大的项目，管理者必须设立规章，规

定数据使用者应如何评估风险、如何规避或者减轻潜在伤害。这将激发数据的创新性再利用，同时也确保个人免受无妄之灾。

对大数据使用进行正规评测及正确引导，可以为数据使用者带来切实的好处：很多情况下，他们无须再取得个人的明确同意，就可以对个人数据进行二次利用。相反地，数据使用者也要为敷衍了事的评测和不达标准的保护措施承担法律责任，诸如强制执行、罚款甚至刑事处罚。数据使用者的责任只有在有强制力规范的情况下才能确保履行到位。

为了了解它是如何运用到实践当中的，我们以第4章提到的数据化个人坐姿信息为例。假设一家公司出售了一项以驾驶员坐姿为特定识别符的汽车防盗技术。然后，它对收集到的信息进行分析，预测驾驶员的注意力状态（比如昏昏欲睡、醉酒以及生气），以此向周围其他驾驶员发出警报以防发生交通事故。根据目前的隐私规范，他可能需要新一轮的告知与许可，因为这样使用信息是未经驾驶员授权的。但是如今，在数据使用者的责任承担体系下，他们就会评估预期用途的危险性。如果发现危害性很小，他们就可以着手实施预定计划并实现提高驾驶安全性的目标。

大数据洞察

将责任从民众转移到数据使用者很有意义，也存在充分的理由，因为数据使用者比任何人都明白他们想要如何利用数据。他们的评估（或者由他们所雇用的专家制定的评估）避免了商业机密的泄露。也许更为重要的是，数据使用者是数据二级应用的最大受益者，所以理所当然应该让他们对自己的行为负责。

此外，与目前大多数隐私保护法所要求的不一样，数据使用者达到了最初目的之后，法律上不再规定必须删除个人信息。相反，数据使用者被允许较长时间地保存数据，虽然不能永远。这是一个意义重大的变革，因为，就像我们所知道的，只有开发数据的潜在价值，对数据价值进行最大程度的挖掘，近代“莫里”们才能发展繁荣，并促进自身和社会的同步进步。总之，社会必须平衡二次运用的优势与过度披露所带来的风险。

为实现这一平衡，监管机制可以决定不同种类的个人数据必须删除的时间。再利用的时间框架则取决于数据内在风险和社会价值观的不同。一些国家也许会更谨慎，而某些种类的数据也许会更敏感。

这一方式通过限制个人信息存储和处理的时间而保护了个人隐私，也可以消除“永久记忆”的恐慌——永不磨灭的数字记录让人无法告别过去。我们的个人数据就像达摩克利斯之剑一样悬在头上，多年之后也会因为一件私事或者一次遗憾的购买记录而被翻出来再次刺痛我们。时间限制也激励数据使用者在有权限的时间内尽力挖掘出数据的价值。这就是我们认为更适用于大数据时代的平衡：公司可以利用数据的时间更长，但相应地必须为其行为承担责任以及负有特定时间之后删除个人数据的义务。

大数据洞察

除了管理上的转变，即从个人许可到数据使用者承担相应责任的转变，我们也需要发明并推行新技术方式来促进隐私保护。一个创新途径就是“差别隐私”：故意将数据模糊处理，促使对大数据库的查询不能显示精确的结果，而只有相近的结果。这就使得挖出特定个人与特定数据点的联系变得难以实现并且耗费巨大。

信息模糊处理听起来似乎破坏了其解读价值，但是也并不一定，至少这是一个折中的好办法。例如，技术政策专家特别提到Facebook将用户信息向潜在广告客户公布就是依靠差别隐私：只能得出大概数量，所以它们不会暴露个人身份。查询亚特兰大州对Ashtanaga瑜伽感兴趣的亚洲女性，会得出诸如“400左右”这样的结果而非一个定值。因此，利用这个信息排查到个人是不可能的。

个人隐私保护从个人许可到数据使用者承担责任的转变是一个本质上的重大变革。我们必须将类似范围内的变革应用到大数据预测中去，以维护人类的自由和责任。

管理变革2：个人动因VS预测分析

在法庭上，个人对自己的行为负有责任。一番公正的审讯之后，审判员会做出公平公正的判决。然而，在大数据时代，关于公正的概念需要重新定义以维护个人动因的想法：人们选择自我行为的自由意志。简单地说，就是个人可以并应该为他们的行为而非倾向负责。

在大数据时代之前，这是明显而基本的自由权利。事实上，明确到不需要进行说明。毕竟，我们的法律体系就是这样运作的：通过评判人们过去的行为使之为其行为承担责任。然而，有了大数据，我们就能预

测人的行为，有时还能十分准确。这诱使我们依据预测的行为而非实际行为对人们进行评定。

大数据洞察

身处大数据时代，我们必须拓宽对公正的理解，必须把对个人动因的保护纳入进来，就像目前我们为程序公正所做的努力一样。如若不然，公正的信念就可能被完全破坏。

通过保证个人动因，我们可以确保政府对我们行为的评判是基于真实行为而非单纯依靠大数据分析。从而，政府只能依法对我们过去的真实行为进行追究，而不可以追究大数据预测到的我们的未来行为；或者，在政府评判我们过去的行为时，也应该防止单纯依赖大数据的分析。例如，在对两家涉嫌操纵价格的公司进行调查时，我们完全可以借助大数据分析先作出大概判定，然后监管机构再以传统手段立案和进行调查。不过，当然不能只因为大数据分析预测它们可能犯罪，就判定其有罪。

相似的原理应当运用到政府领域之外，比如公司在进行关乎我们个人利益的重大决策时——雇用与解雇，同意按揭或者拒发信用卡。如果他们单纯依据大数据预测作出这些决策，特定的防护措施就必须到位。

●第一，公开原则。因为这将直接影响到个人，所以必须公开用来进行预测分析的数据和算法系统。

●第二，公正原则。具备由第三方专家公证的可靠、有效的算法系统。

●第三，可反驳原则。明确提出个人可以对其预测进行反驳的具体方式（这类似于科学研究中披露任何可能影响研究结果的因素的传统）。

●最重要的是，要确保个人动因能防范“数据独裁”的危害——我们赋予数据本不具备的意义和价值。

保护个人责任也同样重要。也许这一点很具有吸引力——社会无论何时做出关乎他人的决策时，都不再需要决策者们承担责任。相反，它会将重心转移到风险管理上，即评测可能性以及对其进行风险评估。有

了所有看似客观的数据，对我们的决策过程去情绪化和去特殊化，以运算法则取代审判员和评价者的主观评价，不再以追究责任的形式表明我们的决策的严肃性，而是将其表述成更“客观”的风险和风险规避，听起来都是不错的主意。

比如说，大数据强烈诱使我们隔离那些被预言将会犯罪的人们，以减少风险的名义对其进行不断审查，即使他们确实在为尚不需承担责任的事情接受惩罚。设想一下，“预测警务”的运算法则鉴定某个青少年在未来五年内很可能犯重罪。结果，当局决定派遣一名社会工作者每月拜访他一次以对其进行监视，并尽力帮助他解决问题。如果该少年及其亲属、朋友、老师或雇主将这种拜访视为一种耻辱（这种情况极有可能发生），那么这就起到了惩罚的作用，的确是对未发生的行为的惩罚。然而，如果这种拜访完全不被视为惩罚，而只是为了减少问题出现的可能，即作为一种将风险降至最低的方式（在这里指的是将破坏公共安全的犯罪风险减到最小），情况照样好不到哪儿去。社会越是用干预、降低风险的方式取代为自己的行为负责，就越会导致个人责任意识的贬值。主张预测的国家是保姆式的国家，而且远不止如此。否认个人为其行为承担责任实际上就是在摧毁人们自由选择行为的权利。

如果国家做出的许多决策都是基于预测以及减少风险的愿望，就不存在所谓个人的选择了，也不用提自主行为的权利。无罪，无清白。如此一来，世界不止不会进步，反而在倒退。

大数据洞察

大数据管理的基本支撑是保证我们依然是通过考虑他人的个人责任对其进行评判，而不是借助“客观”数据处理去决定他们是否违法。只有这样，我们才是把其当作人来对待——当作有行为选择自由和通过自主行为被评判的人。这就是从大数据推论到今天的无罪推定原则。

管理变革3：击碎黑盒子，大数据算法师的崛起

目前，计算机系统做出决策的方式是基于程序明确设定所需遵循的规则。这样，如果它们的决策出错（这是不可避免的），我们就可以回过头来找出计算机做出错误决策的原因。“为什么外部感应器遭遇空气湿度激增的情况时，智能飞行系统使飞机上升了5度？”等。现在的计算机编码能被解码、检查，并且可以解读其决策依据——无论多么复杂，至少对于懂得如何解码的人不存在问题。

然而，有了大数据分析，这种追踪会变得愈发困难。对人们而言，进行预测分析的计算机系统往往过于复杂，根本无法理解。但当计算机按程序设置明确执行一系列指令时，情况就不一样了。例如1954年早期，在IBM将俄文译成英文的翻译程序中，人们就能轻松理解一个单词译成另一个单词的原因。但是，对于谷歌利用几十亿页的翻译数据开发出的翻译系统，当其将英文单词“light”译成“光”而不是“重量轻”时，就不可能清楚地解释如此选择的原因，毕竟这个预测分析是基于海量数据和庞大的统计计算之上的。

大数据的运作是在一个超出我们正常理解的范围之上的。例如，谷歌所确定的与流感相关的检索词条是通过测试了4.5亿个数学模型而得出的。而辛西亚·鲁丁最初为判断沙井盖是否会着火设定了106个预测器，因此才能向联合爱迪生电力公司的经理解释为何程序要求优先检查某个沙井盖。“可解释性”正如在人工智能界所称的一样，对于不仅想知道“是什么”更想知道“为什么”的人类来说非常重要。可是，如果系统自动生成的不是106个预测器，而是601个的话，那该怎么办呢？虽然其中大部分都没有多大用途，但是一旦汇聚起来就能提高模型的准确性，而预测的基础就会变得惊人地复杂。如此的话，辛西亚·鲁丁如何能说服联合爱迪生电力公司的经理再分配它们本就不多的预算呢？

在这些背景下，我们能看到大数据预测、运算法则和数据库有变为黑盒子的风险，这个黑盒子不透明、不可解释、不可追踪，因而我们对其信心全无。为了防止这些情况的出现，大数据将需要被监测并保持透明度，当然还有使这两项得以实现的新型专业技术和机构。它们将为许多领域提供支持，在这些领域里社会需要检测预测结果并能够为被其错误引导的人群提供弥补方法。

社会发展出现过很多这种情况，当一个特定领域变得特别复杂和专门化之后，就会催生出对运用新技术的专门人才的迫切需求。在一个多世纪以前，法律、医学、会计以及工程学领域都经历过这种转型。不久前，计算机安全和隐私顾问的突然兴起，证实了公司都在遵循由一些组织确立的行业最佳做法，如国际标准化组织，它是为满足这个领域对准则的需要而自发形成的。

大数据洞察

大数据将要求一个新的人群来扮演这种角色，也许他们会被称作“算法师”。他们有两种形式：在机构外部工作的独立实体和机构内部

的工作人员——正如公司有内部的会计人员和进行鉴证的外部审计师。

这些新的专业人员会是计算机科学、数学和统计学领域的专家，他们将担任大数据分析和预测的评估专家。他们必须保证公正和保密，就像现在的审计员和其他专业人员所做的一样。他们可以评估数据源的挑选，分析和预测工具的选取，甚至包括运算法则和模型，以及计算结果的解读是否正确合理。一旦出现争议，他们有权考察与分析结果相关的运算法则、统计方法以及数据集。

如果2004年美国国土安全部配备有一名算法师，它也许不会生成一份这么差劲的禁飞名单，竟然把马萨诸塞州参议员特德·肯尼迪都列入了其中。最近在日本、法国、德国和意大利，算法师也可以发挥作用，这些国家的很多人认为谷歌的“自动完成”特征程序诽谤了他们。这是一个生成与姓名相关的普遍搜索词的程序，它很大程度上依据的是之前的搜索频率：这些词条根据数学概率进行排名。如果类似“犯罪”或者“娼妓”这样的字眼出现在你姓名旁边，而碰巧被你的业务伙伴或者爱人看到了，你能不气疯吗？

我们将“算法师”的概念视为是在以市场为导向来解决这些问题，这也就避免了以侵入式的规章来解决问题。他们和20世纪早期为了处理泛滥的财务信息而出现的会计以及审计员一样，都是为了满足新需求而出现的。一般人很难理解这样的数字冲击，所以必须有一群以一种灵活的自我监管方式组织起来的专业人员去保护大众的利益。于是，提供专门的金融监管服务的新公司就这样应运而生。如此一来，这种新类型的专门人才也帮助社会大众增强了他们对经济本身的信心。大数据可以也应该从算法师给予的类似信心提振中获利。

外部算法师

外部算法师将扮演公正的审计员的角色，在客户或政府所要求的任何时候，根据法律指令或规章对大数据的准确程度或者有效性进行鉴定。他们也能为需要技术支持的大数据使用者提供审计服务，还可以为他们证实大数据应用程序的健全性，例如反欺诈技术或者股票交易系统。最后，他们将和政府商议公共领域大数据的最佳使用办法。

就像医学、法律和其他行业一样，我们设想这个新行业会有自己的行业规范。算法师的公正、保密、资历以及专业水准可用严苛的责任规范来进行强制约束；如果他们不能达到这些标准，就可能被起诉。他们

可以调任为审讯中的专家证人，或在审讯中遇到特别复杂的大数据问题时被法官委派为“法院专家”——主要是指某一个学科领域专家为案件审理提供援助。

此外，当人们认为他们受到大数据预测危害——被拒绝手术、被拒绝假释、被拒绝抵押贷款时，便可以向算法师咨询并针对这些决定提起诉讼。

内部算法师

内部算法师在机构内部工作，监督其大数据活动。他们不仅要考虑公司的利益，也要顾及受到公司大数据分析影响的其他人的利益。他们监督大数据的运转，任何认为遭受其公司大数据危害的人都会最初与他们取得联系。在公布大数据分析结果之前，他们也对其完整性和准确度进行审查。为了扮演好这两个角色，算法师首先要做到的就是必须在工作机构内部拥有一定程度的自由和公正。

个人为公司工作却又要保持公正似乎违背常识，但事实上这十分常见。大型金融机构的监管部门是一个例子，许多公司的董事会也是如此，他们是对股东负责而非管理者。许多传媒公司，包括《纽约时报》、《华盛顿邮报》都会雇用外部监察人来维护公众信任。他们解决读者的问题，当他们发现存在不当行为时，也经常公开责难雇主。

甚至，与内部算法师更类似的职业也同样存在，即负责确保企业不滥用个人信息的职业人。例如在德国，具有一定规模的公司（有10个或以上人员处理个人信息）必须任命一名数据保护代表。20世纪70年代以来，数据保护代表们逐渐形成了自己的职业道德和团体精神。他们进行定期会面，分享最好的实践经验并进行培训，他们拥有自己专门的媒体和会议，他们也成功地实现了一方面忠于雇主，另一方面忠于自己作为公证人的职责。德国的企业数据保护代表们取得了很大的成功，既充当了企业数据保护监察人，又将信息保密观念嵌入了整个企业运作过程。我们相信，算法师同样也能做到。

管理变革4：反数据垄断大亨

数据之于信息社会就如燃料之于工业革命，是人们进行创新的力量源泉。没有大量鲜活的数据和健全的服务市场，这些创新就实现不了。

在这一章节，我们已经提及了管理上的三个基本转变。随着这些转变的完成，我们相信，大数据的不利影响将会得到控制。然而，随着尚未成熟的大数据产业的不断发展，另一个重要的挑战将会是如何保护极具竞争力的大数据市场。我们必须防止21世纪数据大亨的崛起，它相当于19世纪强盗大亨的现代翻版，那些强盗大亨曾垄断了美国的铁路、钢铁生产和电报网络。

为了管理这些新兴行业，美国制定了适应性极强的反垄断条例。最初是在19世纪为铁路行业制定的，后来又被应用到了掌管商业信息的其他公司，从20世纪最初十年的国家收银机公司（National Cash Register），到20世纪60年代的IBM、70年代的施乐公司、80年代的AT＆T、90年代的微软和今天的谷歌。这些公司所开辟的技术成了经济结构中“信息基础设施”的核心组成部分，所以为了防止它们垄断，法律的支持必不可少。

为了确保给大数据提供一个与早期技术领域情况相当的活跃的市场环境，我们应该实现数据交易，比如通过授权和协同合作的方式。但是，这就引发了一个问题：精心达到平衡的数据独有权，是否能让社会大众从中获利？虽然听起来有点挑衅的意味，但是这是否能像知识产权一样有利于社会呢？诚然，要达到这样的效果，对于决策者来说，是一个艰难的任务；而对于普通人来说，则充满了风险。技术发展变幻莫测，无从定论，大数据也无法预测自己的未来。监管人员需要既大胆又细心，而实现这两者的平衡，可以学习反垄断法的发展历史。

反垄断法遏制了权力的滥用。然而令人惊奇的是，这些条例能从一个领域完美转移到另外一个领域，并且适用于不同类型的网络产业。这种不带任何偏袒的强有力的规章非常实用，因为它提供的是一个平等的竞争平台，一开始便没有任何优劣之分。因此，为了促进大数据平台上的良性竞争，政府必须运用反垄断条例。而且，就像世界上一些大型的数据拥有者那样，政府也应该公布其数据。令人高兴的是，这一切正在发生。

反垄断法的经验是，一旦确定了极重要的原则，管理者就要将之付诸行动，以确保保护措施的实施到位。同样，我们提出了三项策略，包括隐私保护从个人许可到数据使用者承担责任的转变，在使用预测分析时考虑个人动因以及催生大数据审计员，也就是算法师。这都将是大数据时代对信息进行有效、公正管理的基础。

伴随着从核技术到生物工程学其他领域的发展，人类总是先创造出可能危害自身的工具，然后才着手建立保护自己、防范危险的安全机制。在这方面，大数据也和其他领域的新技术一样，带来了无法彻底解决的问题。另外，它们也不断对我们管理世界的方法提出挑战。而我们的任务是要意识到新技术的风险，促进其发展，然后斩获成果。

正如印刷机的发明引发了社会自我管理的变革，大数据也是如此。它迫使我们借助新方式来应对长期存在的挑战，并且通过借鉴基本原理对新的隐患进行应对。不过，推进科学技术进步的同时，应确保人类自身的安全。因此，我们不能让大数据的发展超出我们可以控制的范围。

结语 正在发生的未来

大数据并不是一个充斥着算法和机器的冰冷世界，人类的作用依然无法被完全替代。大数据为我们提供的不是最终答案，只是参考答案，帮助是暂时的，而更好的方法和答案还在不久的未来。

凡是过去，皆为序曲

麦克·弗劳尔（Mike　Flower）是21世纪初曼哈顿地区检察官办公室的一名律师，负责过从谋杀案到华尔街金融犯罪等各式各样的诉讼案件，后来他转到一家大型的企业律师事务所工作。在办公桌后度过了无聊的一年后，他决定离开。他想做些更有意义的事情，随即想到了去帮助重建伊拉克。在公司的一位朋友给高层打了几个电话后，弗劳尔被派去了绿色区域，也就是美军驻巴格达市中心的安全地带，成为萨达姆·侯赛因审判律师团中的一名律师。

他主要负责后勤事务，而不是相关的法律工作。他负责将证人运送到绿色区域，其间需要安全通过无数每天都会上演的简易爆炸装置袭击（IED）。他看到了军队人员是如何将这当作数据问题来进行处理的。情报分析员结合实地考察报告和过去IED袭击地点、时间和人员伤亡的详细信息，据此预测一天中最安全的运送路线。

在弗劳尔回到纽约两年后，他意识到这些方法其实是一个打击犯罪的有力方式——比他过去作为检察官所掌握的方式更棒。弗劳尔之后被任命为专案组成员，研究可能揭露2009年次贷丑闻罪犯的数据。这个团队做得非常出色，以至于一年后，纽约市长布隆伯格要求扩大规模。弗劳尔成了全市首个“分析主任”，他的任务就是找到最优秀的数据科学家并组建团队，利用城市尚未开发的信息库，收获一切可能的效益。

弗劳尔为了找到合适的人而广泛撒网。“我对经验丰富的统计学家没有兴趣，我担心他们不愿意采取这种新方法来解决问题。”当他采访统计学家对金融诈骗项目的看法时，他们往往会提出晦涩难懂的数学问题。“我甚至没有想到我要使用什么样的模式。我想要可执行的洞察力，这是我所关注的。”他说。最后，弗劳尔一共挑选了5个人组成团队，他称他们为“小伙子”。除一名成员外，其他都是刚毕业一两年的经济学专业学生，而且从未在大城市生活过，但他们都很有创造力。

他们最早处理的事件之一是“非法改建”，即将一套住房隔出很多小房间，这样就能够多容纳10倍的人。非法改建会带来巨大的火灾隐患，也是犯罪、毒品、疾病和虫害孵化的温床。乱麻一般的分机线绳会沿墙壁穿过，电炉可能会放在床单的上面，一旦发生火灾，人也许会被裹得紧紧地葬身火海。2005年，两名消防队员因营救非法改建住房的人而死

亡。纽约市每年会受到约25000起非法改建的投诉，但只有200名检察员在处理这些事情。似乎没有什么好办法来区分简单的滋扰事件和严重的爆炸起火事件。但对弗劳尔和他的小伙子们来说，这看起来更像是一个可以用大量数据来解决的问题。

他们将城市里的90万栋建筑都列在表上，然后输入来自19个不同机构的数据集。这些数据显示了建筑业主是否拖欠了应缴房产税，是否有止赎诉讼，是否有公用设施使用异常或导致服务消减的未付款项。他们还输入了建筑类型、修建时间、救护车访问次数、犯罪率和啮齿动物投诉等信息。然后，他们将这些数据与五年来的火灾严重性排名数据进行对比并得到一个模型，以此预测哪些投诉迫切需要调查。

最初，许多数据形式都不可用。例如，在一个城市里，描述地理位置的方法不是唯一的，每个机构和部门似乎都有自己的描述方式。建筑部门给予每个建筑物一个独特的号码；房屋维护部门也有自己独有的编号系统；税务部门依照街区和地皮，给予每个建筑物特定的标识；警察局采用笛卡尔坐标系；消防局依托“电话亭”临近体系，将建筑物与各个消防站的位置联系在一起，尽管这些电话亭并非真实存在。弗劳尔的小伙子们处理这种不统一的方式是：以笛卡尔坐标系为基础，取用建筑物周围的一片辐射范围并从其他机构的数据库调取地理位置数据，从而建立一个系统。这些数据本身并不精确，但是巨大的信息量弥补了这点瑕疵。

尽管如此，他们并不满足于仅仅对数据进行运算，而是会到现场观看检查员的工作。他们不断做着大量笔记，并询问一切流程的开展效果。当一个头发斑白的领头人哼了一声说“找到那个建筑不是问题”时，他们很想知道为什么这个人会这么自信。但领头人自己也说不清楚为什么，不过弗劳尔的小伙子们渐渐发现，这种直觉来自建筑物外新的砖工，它暗示着建筑物的主人很重视这个地方。

小伙子们回到自己的工作间，钻研着如何能将“新的砖工”作为一种信号融入到他们的模型中，毕竟，砖块是没有被数据化的。但是可以肯定的是，做任何外部砖工都需要城市许可证。这些信息都可以用于提高系统的预测功能，并且他们发现，很多传统意义上可疑的特点其实都无关紧要。

这种分析法或许揭示了：有些历史最悠久的做事方法并不是最好的，就好比《点球成金》中的球探们不得不接受他们直觉中的缺陷一

样。例如，人们将城市“311”投诉热线的来电数量作为衡量问题严重性的指标，来电越多说明问题越严重。但是这种引导是错误的。在繁华的上东区发现一只老鼠也许会在仅仅一个小时之内引发30个投诉电话。然而在布朗克斯区，街坊只有在看到成群结队的老鼠时，才会觉得有必要打个投诉电话。同样，很多非法改建的投诉也许会让人们议论纷纷，但是其后果并没有那么严重。

2011年6月，弗劳尔和他的小伙子们开始启用他们的系统和方法。他们每周浏览一次可归为“非法改建”一类的投诉，将他们认为前5%有火灾危险的投诉转交给检查员立刻跟进。当拿回结果时，所有人都惊呆了。

大数据的力量

在大数据分析之前，检查员会先跟进他们认为最急迫的投诉，而只有13%的案件足够严重，需要立刻去处理。现在，他们立即处理的投诉案件占他们所有安全监测的70%。大数据节省了检查员的时间，将效率提高到原来的5倍。他们的工作也越来越令人满意：精力都集中于最严重的问题。他们新发现的成果还带来了额外利益。非法改建中的火灾更可能导致消防员受伤或死亡，概率是普通案件的15倍。消防局因此非常满意。弗劳尔和他的小伙子们就好像巫师一样，手中的水晶球让他们可以预见未来，看到哪里是最危险的。他们利用了大量搁置多年的数据，这些数据自收集以来几乎没被用过。他们用新的方法管理这些信息，从而提取出它们真正的价值。他们从大的信息库中释放了洞察力，而这在较小数据中是做不到的，这就是大数据的缩影。

纽约市分析炼金师的经验凸显了本书中的不少主题。他们使用了庞大的数据量，而不仅是一些数据。他们所列的城市建筑基本上可以视为“样本=总体”。位置信息或救护车记录等数据比较凌乱，但是这并没让他们就此放弃。更多数据所带来的好处远比原始信息少所带来的弊端更重要。他们之所以能取得成功，是因为城市的很多功能都以数据的形式呈现（尽管存在不一致），从而使他们能够处理和使用这些信息来提高预测效果。

专家暗示，无论是自大的统计学家还是专管投诉热线的公务员，在数据驱动方法面前都应退居次席。与此同时，弗劳尔和他的小伙子们不断地让经验丰富的检查员来测试他们的模型，借鉴检察员们的经验，使系统表现得更好。这个项目成功最重要的原因是，它更多依赖的是相关

关系而非因果关系。

“我对因果关系不感兴趣，除非它用行动说话。”弗劳尔解释道。“因果关系是别人的事，坦白说，谈论因果关系是非常冒险的。我不认为有人提出房产止赎程序和那个地方是否长期存在结构性的火灾风险之间有任何关系。我认为这么想很愚蠢。他们会认为有一些潜在的因素，但没有人会站出来承认。我不想深究这个，我需要一个能够把握的特定数据点来告诉我它的意义。如果它很重要，我们就会采取行动。如果不重要，我们就不会行动。你知道，我们有真正需要解决的问题。我不会闲逛，或者像现在一样想着因果关系的事儿。”

大数据时代，名副其实的“信息社会”

大数据在实用层面的影响很广泛，解决了大量的日常问题。大数据更是利害攸关的，它将重塑我们的生活、工作和思维方式。在某些方面，我们面临着一个僵局，比其他划时代创新引起的社会信息范围和规模急剧扩大所带来的影响更大。我们脚下的地面正在移动。过去确定无疑的事情正在受到质疑。大数据需要人们重新讨论决策、命运和正义的性质。我们的世界观正受到相关性优势的挑战。拥有知识曾意味着掌握过去，现在则更意味着能够预测未来。

当我们准备开发电子商务、寓生活于互联网、进入计算机时代或者拿起算盘时，这些事情比那些代表他们的问题更加重要。我们寻找原因的想法可能被高估了，很多情况下，弄清楚“是什么”比找寻“为什么”更加重要，因为前者表明事实才是我们生活和思维的基础。这些问题可能没有答案。或许，它们是关于人在宇宙中的位置以及能否在喧嚣混乱、不可理喻的世界中寻找到意义这一永恒争论的一部分。[1]

最终，大数据标志着“信息社会”终于名副其实。我们收集的所有数字信息现在都可以用新的方式加以利用。我们可以尝试新的事物并开启新的价值形式。但是，这需要一种新的思维方式，并将挑战我们的社会机构，甚至挑战我们的认同感。可以肯定的是，数据量将继续增长，处理这一切的能力也是如此。但是，现在大多数人都认为大数据是一个技术问题，应侧重于硬件或软件，而我们认为应当更多地考虑当数据说话时会发生什么。

大数据洞察

现在，我们可以获得比以前更多的信息并进行分析。在我们诠释世界时，数据不再是限制我们努力的因素了。我们可以利用更多的数据，某些情况下，甚至是全部数据。但是这需要我们采取非传统的方法，特别是要改变我们理想中构成有用信息的因素。

除了纠结于数据的准确性、正确性、纯洁度和严格度之外，我们也应该容许一些不精确的存在。数据不可能是完全对或完全错的。当数据的规模以数量级增加时，这些混乱也就算不上问题了。事实上，它甚至可以是有好处的，因为当我们只想使用一小部分时，无须捕捉这么多的知识细节。又因为我们可以用更快更便宜的方式找到数据的相关性，并

且效果往往更好，而不必努力去寻找因果关系。当然在某些情况下，我们仍然需要精心策划的数据来做因果关系研究和控制实验，如测试药物的副作用或设计关键的飞机部件。但是在日常情况下，知道“是什么”就已经足够，不必非要弄清楚“为什么”。大数据的相关性将人们指向了比探讨因果关系更有前景的领域。

这些相关性能让我们节省机票钱和预测流感爆发，并知道在一个资源有限的世界中应该检查哪些沙井盖和过度拥挤的建筑物。它可以帮助健康保险公司不做体检就能决定保险覆盖面，并降低提醒病人服药的成本。通过大数据的相关性，语言可以得到翻译，汽车可以在预测的基础上自行驾驶。沃尔玛可以了解飓风前应在门店准备哪种口味的蛋挞[2]。当然，如果能从中得到因果关系更好。问题是，因果关系往往很难找到，通常我们认为找到了的时候，都是在自欺欺人。

我们之所以能做所有这些事，新工具只是个很小的因素，无论是更快的处理器、更多的存储器，还是更智能的软件和算法。这些固然重要，但是更为根本的原因是我们拥有了更多的数据，继而世界上更多的事物被数据化了。诚然，人类量化世界的雄心先于计算机革命，但是数字工具将数据化提升到了新的高度。不仅移动电话能够跟踪到我们呼叫的人和我们所在的位置，而且同样的数据也能用于断定我们是否生病了。不久之后，它或许还能够辨别我们是否恋爱了。

大数据洞察

我们“做新、做多、做好、做快”的能力能释放出无限价值，产生新的赢家和输家。大部分的信息价值来自二级用途，即潜在价值，而不是我们所习惯认为的基本用途。结果，对于大多数数据来说，尽可能多地收集、等待信息增值并且让其他更适合挖掘其价值的人来分析它才是明智之举（前提是此人能够分享开发出的利润）。

能置身于信息流中央并且能收集数据的公司通常会繁荣兴旺。有效利用大数据需要专业技术和丰富的想象力，即一个能容纳大数据的心态，但价值的核心归功于数据本身。有时，重要的资产并不仅仅是能清楚看到的信息，更是从人们与信息交互中收集到的数据废气，聪明的公司可以用它来改善现有的服务，或推出全新的服务。

大数据同时也给我们带来了巨大的风险。它使得目前用以保护隐私的法律手段和核心技术失去了效果。过去个人身份信息包含的是名字、

社会安全号码、税收记录等，其构成简单明了。因此隐私保护相对比较简单，只要确保不使用这些信息即可。而今天，即使是最无害的数据，只要被数据收集器采集到足够的量，也会暴露出个人身份。匿名化或是单纯隐藏已不再适用。不仅如此，现在要是对某人进行监督，必定会侵犯到较之以往范围更广的个人隐私内容。因为政府在管理上不仅要求个人信息尽可能完善，还记录了其所有的社会关系、交往和交流信息。

无论大数据如何威胁到隐私保护，最让人们头疼的都是行为倾向问题。大数据预测的准确性越来越高，它能够预测行为的发生，在人们犯错之前，提前惩处。因为预测的结果几乎不可反驳，人们也就无法为自己开脱。但这种基于预测得出的惩罚不仅违背自由意志的原则，同时也否定了人们会突然改变选择的可能性（无论可能性有多小）。当我们给一个人判定责任（并给予惩罚）时，必须牢记人类意志的神圣不可侵犯性。人类的未来必须保留部分空间，允许我们按照自己的愿望进行塑造。否则，大数据将会扭曲人类最本质的东西，即理性思维和自由选择。

应对大数据的汹涌来袭，我们没有万无一失的方法，必须建立规范自身的新准则。随着社会越来越熟悉大数据的特征和缺陷，我们可以改变一系列的惯例来帮助社会应对这种冲击。我们需要把进行隐私保护的责任从个人转移到数据使用者身上，也就是说，数据使用者应该以负责任的态度使用数据。

在一个预测的时代里，人类的自由意志神圣而不可侵犯，这一点不可轻视。我们不仅需要承认个人进行道德选择的能力，还要强调个人应为自我行为承担责任。社会则必须采取新的保护措施：接受一种新的职业人，也就是数据算法师，对大数据进行深度分析。如此，因为大数据而变得可预测的世界，才不会陷入一个用一种未知取代另一种未知的困境中，不会变成一个黑匣子。

大数据将成为理解和解决当今许多紧迫的全球问题所不可或缺的重要工具。例如要应对气候变化问题时，需要对污染相关数据进行分析，得出最佳方案，来指导努力方向，找出缓解问题的方法。全球范围内遍布的大量传感设备，包括智能手机内部的传感器，使我们能够以更高的细节水平模拟环境。而世界贫困人口迫切需要提高医疗保健服务，降低医疗费用,这很大程度上可以靠自动化来实现。当下许多似乎需要人类判断才能进行的事情，其实完全可以交由电脑来做，比如癌细胞活检、传染病爆发前期的模式预测等。

大数据也被用于发展经济和理解如何预防冲突。基于手机动向数据显示，非洲许多贫民窟地区经济活动十分活跃。大数据还揭示了最可能引发种族关系紧张的社区以及解除难民危机的方式。只有当科技应用至生活的方方面面时，大数据的使用范围才能进一步扩大。

大数据能帮助我们更好地进行已有的工作，并处理全新的事务。但它绝不是魔术棒，不会带来世界和平，无法根绝贫穷问题，更不能创造出另一个毕加索。大数据不能造婴儿，虽然它确实可以救助早产儿。不要多久，我们将在生活的各个方面使用到大数据，如果不用的话还可能会引起些许焦虑，这种情况就像普通体检查不出问题时，会希望有医生帮我们预约X光进行检查。

当大数据成为日常生活的一部分后，它将会极大地改变我们对未来的看法。大约五百年前，欧洲在逐渐发展为更加自由、科学、文明的世界的进程中，欧洲人经历了对时间认知的重大转变。在此之前，时间被认为是循环的，生命也是轮转的。每天或每年与过去的日子如出一辙，甚至连生命的终结也与起点相似，因为濒死的成人会显示出孩子的特征。认知转变后，时间变作线性的，成了一条岁月演变过程，过程中世界因人变化，生命的轨迹也受到相应的影响。如果说这以前的历史中，过去、当下、未来的概念是完全交织在一起的，那么通过塑造当下，人类现在便有了过去可以回顾，有了未来可以展望。

虽然我们可以塑造当下，但未来却从过去的“完全可预测”转变为一块开放又原始、广阔而空白的帆布，所有人都可以在上面依据自己的价值，努力裁剪塑形。“现代”的一个定义性特征便是人类感到自己是命运的主人，这使我们与生活在宿命论桎梏中的先辈们截然不同。但是大数据预测却又使我们的生命帆布不再那么开放、原始和纯净。对于善于运用科技解读未来的人来说，我们的未来不再是只字未书的画布，而是似乎已经着上了淡淡的墨痕。未来的可预知性似乎缩小了塑造命运的空间。潜在的可能性在概率的圣坛上被解剖。[3]

与此同时，大数据又意味着我们将永远受困于过去的行为，这些行为在预知我们下一步动作的预测过程中与我们作对，即我们永远无法逃避已发生的事。莎士比亚曾写道：“凡是过去，皆为序曲。”大数据通过运算将这句话铭刻，无论结果好坏——无论这句话是否会浇熄我们迎接下一个日出的热情，是否会打击我们留名于世的渴望。

其实，事实很有可能是相反的。知道行为在未来如何谢幕，我们便

可以采取补救措施，避免问题发生并改善结局。我们能在期末考试之前早早发现有退步趋势的学生。我们能检测到微小的癌变，赶在疾病完全爆发前根治。我们能看到青春期意外妊娠的可能性，或是预测到某种犯罪生涯，然后尽力干预，避免出现可能的悲剧结局。例如拥挤的纽约住宅着火的时候，如果能事先知道并从几间最可能是火源的公寓着手，将会免除一场致命的火灾。

没有什么是上天注定的，因为我们总能就手中的信息制定出相应的对策。大数据预测结果也并非铁定，而只是提供了一种可能性，也就是说，只要我们愿意，结局可以改写。我们可以判断出迎接未来的最佳方式，摇身变作未来的主人，正如莫里在海与风的广阔世界中乘风破浪一般。在过程中我们无须理解宇宙的奥秘或是去证明神的存在，因为大数据已经帮我们做好了。

---

[1] 当有些大数据拥护者叫嚣“理论已死”的时候，各位读者是否回到了一个多世纪前让尼采喊出“上帝已死”的时代？大数据恐怕会让我们在寻找意义塑造价值的道路上面对比尼采当时更困难的境况。——译者注

[2] 答案：草莓味。——作者注

[3] 意指潜在可能发生的所有事情将借由大数据分析而获得，一种概率描述。——译者注

更大的数据源于人本身

大数据改造了我们的生活，它能优化、提高、高效化并最终捕捉住利益，那直觉、信仰、不确定性和创意还能扮演什么角色呢？

就算大数据无法教会我们所有事情，只要能帮助我们表现更佳、更富效率、取得进步，就算缺乏深入理解也是很有用的了。一贯如是地坚持下去才有效力。即使你不明白为什么付出的努力得不到回报，但相比不努力，你要明白你已经在改善事情的结局了。纽约的弗劳尔和他的“小伙子们”也许并没有圣人圣明的判断力，但他们确实在拯救生命。大数据不会即刻提高效率，但经受住时间的考验后，它将生出智慧的结晶。

大数据洞察

大数据并不是一个充斥着运算法则和机器的冰冷世界，其中仍需要人类扮演重要角色。人类独有的弱点、错觉、错误都是十分必要的，因为这些特性的另一头牵着的是人类的创造力、直觉和天赋。偶尔也会带来屈辱或固执的同样混乱的大脑运作，也能带来成功，或在偶然间促成我们的伟大。这提示我们应该乐于接受类似的不准确，因为不准确正是我们之所以为人的特征之一。就好像我们学习处理混乱数据一样，因为这些数据服务的是更加广大的目标。毕竟混乱构成了世界的本质，也构成了人脑的本质，而无论是世界的混乱还是人脑的混乱，学会接受和应用它们才能得益。

在这个利用数据做出决定的世界里，人类存在的目的是什么？难道是为了运用直觉和违背事实？如果所有人都诉诸数据，都利用工具，那时人类的无法预测性即直觉、冒险精神、意外和错误等，反倒可能发挥出重大作用。

如果真变成这样，为人类开辟出一块领地，为直觉、常识和意外运气腾出空间就十分必要，以确保它们不被数据和机器回答挤兑出去。人类最伟大之处正是运算法和硅片没有揭示也无法揭示的东西，因为数据也无法捕捉到这些。并不是“人类最伟大的东西是什么”，而是“什么不是人类最伟大的产物”——真空、人行道上的裂缝、未说出口的话还是未想到的事？

这为“社会进步”的概念提供了重要启示。大数据让我们试验的速度更快，发现的线索更多。这理应能够产生更多的创新成果，但发明的火花却往往存在于数据未显示出的信息之中，因为它并非真实存在，是多大量的数据都永远无法确定或证实的。如果亨利·福特问大数据他的顾客想要的是什么，大数据将会回答，“一匹更快的马。”[1]在大数据的世界中，包括创意、直觉、冒险精神和知识野心在内的人类特性的培养显得尤为重要，因为进步正是源自我们的独创性。[2]

大数据是一种资源，也是一种工具。它告知信息但不解释信息。它指导人们去理解，但有时也会引起误解，这取决于是否被正确使用。大数据的力量是那么耀眼，我们必须避免被它的光芒诱惑，并善于发现它固有的瑕疵。

科技再先进也无法将世界上数据的总量（即最终的样本=总体）尽数收集、储存和加工。例如，欧洲粒子物理研究所（CERN）位于日内瓦的粒子物理实验室在试验中只能收集到不到0.1%的反馈信息，其余信息将同潜在的知识一起消失在乙醚中。这种情况司空见惯。从罗盘和六分仪，到望远镜和雷达，再到今天的全球定位系统，人们总是受到现有测量和认知工具的局限。我们明天使用的工具很可能比今天的强大数倍甚至上千倍，我们现在所拥有的知识较之明天可能就显得微不足道了。要不了多久，当我们回看当今的大数据世界时，就像在看阿波罗11号上仅4Kb内存的导航控制计算机一样，会觉得十分奇特。

我们能收集和处理的数据只是世界上极其微小的一部分。这些信息不过是现实的投影——柏拉图洞穴上的阴影罢了。因为我们无法获得完美的信息，所以做出的预测本身就不可靠。但这也不代表预测就一定是错的，只是永远不能做到完善。这也并未否定大数据的判断，而只是让大数据发挥出了应有的作用。大数据提供的不是最终答案，只是参考答案，为我们提供暂时的帮助，以便等待更好的方法和答案出现。这也提醒我们在使用这个工具的时候，应当怀有谦恭之心，铭记人性之本。

---

[1] 改编自亨利·福特的名言——“如果我当年去问顾客他们想要什么，他们肯定会告诉我：“一匹更快的马。”——作者注

[2] 西方谚语有云：“预测未来最好的办法就是创造未来。”这句话在大数据时代亦应当铭记。在福特时代，任何人都无法从数据中看到汽车将替代马车，福特所创造的是无法预测的全新篇章。——译者注

参考文献

Alter，Alexandra.“Your E-Book Is Reading You.”*Wall Street Journal*，June 29，2012（http：//online.wsj.com/article/SB10001424052702304870304577490950051438304.html）.

Anderson，Benedict.*Imagined Communities*，New Edition.Verso，2006.

Anderson，Chris.“The End of Theory.”*Wired* 16，issue 7（July 2008（http：//www.wired.com/science/discoveries/magazine/16-07/pb_theory）.

Asur，Sitaram，and Bernardo A.Huberman.“Predicting the Future with Social Media.”*Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*，pp.492-499.（An online version is available at http：//www.hpl.hp.com/research/scl/papers/socialmedia/socialm edia.pdf.）

Babbie，Earl.*Practice of Social Research*，12th ed.2010.

Backstrom，Lars，Cynthia Dwork，and Jon Kleinberg.“Wherefore Art Thou R3579X? Anonymized Social Networks，Hidden Patterns，and Structural Steganography.”*Communications of the ACM*，December 2011，pp.133-141.

Bakos，Yannis，and Erik Brynjolfsson.“Bundling Information Goods：Pricing，Profits，and Efficiency.”*Management Science* 45（December 1999），pp.1613-30.

Banko，Michele，and Eric Brill.“Scaling to Very Very Large Corpora for Natural Language Disambiguation.”Microsoft Research，2001，p.3（http：//acl.ldc.upenn.edu/P/P01/P01-1005.pdf）.

Barbaro，Michael，and Tom Zeller Jr.“A Face Is Exposed for AOL Searcher No.4417749.”*New York Times*，August 9，

2006（http：//www.nytimes.com/2006/08/09/technology/09aol.html）.

Barnes，Brooks.“A Year of Disappointment at the Movie Box Office,”*New York Times*，December 25，2011（http：//www.nytimes.com/2011/12/26/business/media/a-year-of-disappointment-for-hollywood.html）.

Beaty，Janice.*Seeker of Seaways*：*A Life of Matthew Fontaine Maury*，*Pioneer Oceanographer*.Pantheon Books，1966.

Berger，Adam L.，et al.“The Candide System for Machine Translation.”*Proceedings of the 1994 ARPA Workshop on Human Language Technology*（1994）（http：//aclweb.org/anthology-new/H/H94/H94-1100.pdf）.

Berk，Richard.“The Role of Race in Forecasts of Violent Crime.”*Race and Social Problems* 1（2009），pp.231-242.

Black，Edwin.*IBM and the Holocaust*.Crown，2003.

boyd，danah，and Kate Crawford.“Six Provocations for Big Data.”Research paper presented at Oxford Internet Institute's “A Decade in Internet Time：Symposium on the Dynamics of the Internet and Society,”September 21，2011（http：//ssrn.com/abstract=1926431）.

Brown，Brad，Michael Chui，and James Manyika.“Are You Ready for the Era of ‘Big Data’?”*McKinsey Quarterly*，October 2011，p.10.

Brynjolfsson，Erik，Andrew McAfee，Michael Sorell，and Feng Zhu.“Scale Without Mass：Business Process Replication and Industry Dynamics.”HBS working paper，September 2006（http：//www.hbs.edu/research/pdf/07-016.pdf；also http：//hbswk.hbs.edu/item/5532.html）.

Brynjolfsson，Erik，Lorin Hitt，and Heekyung Kim.“Strength in Numbers：How Does Data-Driven Decisionmaking Affect Firm Performance?”*ICIS 2011 Proceedings*，Paper 13（http：//aisel.aisnet.org/icis2011/proceedings/economicvalueIS /13；

also available at http：//papers.ssrn.com/sol3/papers.cfm?abstract_id=1819486）.

Byrne，John.*The Whiz Kids*（Doubleday，1993）.

Cate，Fred H.“The Failure of Fair Information Practice Principles.”In Jane K.Winn，ed.，*Consumer Protection in the Age of the “Information Economy”*（Ashgate，2006），p.341 et seq.

Chin，A.，and A.Klinefelter.“Differential Privacy as a Response to the Reidentification Threat：The Facebook Advertiser Case Study.”90 *North Carolina Law Review* 1417（2012）.

Crosby，Alfred.*The Measure of Reality*：*Quantification and Western Society*，1250-1600.Cambridge University Press，1997.

Cukier，Kenneth.“Data，Data Everywhere.”*The Economist* Special Report，February 27，2010，pp.1-14.

“Tracking Social Media：The Mood of the Market.”*The Economist* online，June 28，2012（http：//www.economist.com/blogs/graphicdetail/2012/06/trackin g-social-media）.

Davenport，Thomas H.，Paul Barth，and Randy Bean.“How ‘Big Data’Is Different.”*Sloan Review*，July 30，2012（http：//sloanreview.mit.edu/the-magazine/2012-fall/54104/how-big-data-is-different/）.

Di Quinzio，Melanie，and Anne McCarthy.“Rabies Risk Among Travellers.”*CMAJ* 178，no.5（2008），p.o567.

Drogin，Marc.*Anathema*! *Medieval Scribes and the History of Book Curses*（Allanheld and Schram，1983）.

Dugas，A.℉.，et al.“Google Flu Trends：Correlation with Emergency Department Influenza Rates and Crowding Metrics.”CID Advanced Access，January 8，2012.DOI 10.1093/cid/cir883.

Duggan，Mark，and Steven D.Levitt.“Winning Isn't Everything：Corruption in Sumo Wrestling.”*American Economic Review* 92（2002），pp.1594-1605（http：//pricetheory.uchicago.edu/levitt/Papers/DugganLevitt2002.pdf）.

Duhigg，Charles.*The Power of Habit*：*Why We Do What We Do in Life and Business*.Random House，2012.

Duhigg，Charles.“How Companies Learn Your Secrets.”*New York Times*，February 16，2012（http：//www.nytimes.com/2012/02/19/magazine/shopping-habits.html）.

Dwork，Cynthia.“A Firm Foundation for Private Data Analysis.”*Communications of the ACM*，January 2011，pp.86-95（http：//dl.acm.org/citation.cfm?id=1866739.1866758）.

Economist，The.“Rolls-Royce：Britain's Lonely High-Flier.”*The Economist*，January 8，2009（http：//www.economist.com/node/12887368）.

“Building with Big Data：The Data Revolution Is Changing the Landscape of Business.”*The Economist*，May 26，2011（http：//www.economist.com/node/18741392/）.

“Official Statistics：Don't Lie to Me，Argentina.”The Economist，February 25，2012（http：//www.economist.com/node/21548242）.

“Counting Every Moment.”*The Economist*，March 3，2012（http：//www.economist.com/node/21548493）.

“Vehicle Data Recorders：Watching Your Driving.”*The Economist*，June 23，2012（http：//www.economist.com/node/21557309）.

Edwards，Douglas.*I'm Feeling Lucky*：*The Confessions of Google Employee Number 59*.Houghton Mifflin Harcourt，2011.

Ehrenberg，Rachel.“Predicting the Next Deadly Manhole Explosion.”*WIRED*，July 7，2010（http：//www.wired.com/wiredscience/2010/07/manhole-explosions）.

Eisenstein，Elizabeth L.*The Printing Revolution in Early Modern Europe*.Cambridge：Canto/Cambridge University Press，1993.

Etzioni，Oren，C.A.Knoblock，R.Tuchinda，and A.Yates.“To Buy or Not to Buy：Mining Airfare Data to Minimize Ticket Purchase Price.”SIGKDD '03，August 24-27，2003（http：//knight.cis.temple.edu/~yates//papers/hamlet-kdd03.pdf）.

Frei，Patrizia，et al.“Use of Mobile Phones and Risk of Brain Tumours：Update of Danish Cohort Study.”*BMJ* 2011，343（http：//www.bmj.com/content/343/bmj.d6387）.

Furnas，Alexander.“Homeland Security's ‘Pre-Crime’ Screening Will Never Work.”The Atlantic Online，April 17，2012（http：//www.theatlantic.com/technology/archive/2012/04/homela nd-securitys-pre-crime-screening-will-never-work/255971/）.

Garton Ash，Timothy.*The File*.Atlantic Books，2008.

Geron，Tomio.“Twitter's Dick Costolo：Twitter Mobile Ad Revenue Beats Desktop on Some Days.”*Forbes*，June 6，2012（http：//www.forbes.com/sites/tomiogeron/2012/06/06/Twitters-dick-costolo-mobile-ad-revenue-beats-desktop-on-some-days/）.

Ginsburg，Jeremy，et al.“Detecting Influenza Epidemics Using Search Engine Query Data.”*Nature* 457（2009），pp.1012-14（http：//www.nature.com/nature/journal/v457/n7232/full/nature0 7634.html）.

Golder，Scott A.，and Michael W.Macy.“Diurnal and Seasonal Mood Vary with Work，Sleep，and Daylength Across Diverse Cultures.”*Science* 333（September 30，2011），pp.1878-81.

Golle，Philippe.“Revisiting the Uniqueness of Simple Demographics in the US Population.”*Association for Computing Machinery Workshop on Privacy in Electronic Society* 5（2006），pp.77-80.

Goo，Sara Kehaulani.“Sen.Kennedy Flagged by No-Fly List.”*Washington Post*，August 20，2004，p.A01（http：//www.washingtonpost.com/wp-dyn/articles/A17073-2004Aug19.html）.

Haeberlen，A.，et al.“Differential Privacy Under Fire.”In SEC'11：*Proceedings of the 20th USENIX conference on Security*，p.33（http：//www.cis.upenn.edu/~ahae/papers/fuzz-sec2011.pdf）.

Halberstam，David.*The Reckoning*.William Morrow，1986.

Haldane，J.B.S.“On Being the Right Size.”*Harper's Magazine*，March 1926（http：//harpers.org/archive/1926/03/on-being-the-right-size/）.

Halevy，Alon，Peter Norvig，and Fernando Pereira.“The Unreasonable Effectiveness of Data.”*IEEE Intelligent Systems*，March/April 2009，pp.o8-12.

Harcourt，Bernard E.*Against Prediction*：*Profiling*，*Policing*，*and Punishing in an Actuarial Age*.University of Chicago Press，2006.

Hardy，Quentin.“Bizarre Insights from Big Data.”NYTimes.com，March 28，2012（http：//bits.blogs.nytimes.com/2012/03/28/bizarre-insights-from-big-data/）.

Hays，Constance L.“What Wal-Mart Knows About Customers' Habits.”*New York Times*，November 14，2004（http：//www.nytimes.com/2004/11/14/business/yourmoney/14wal.html）.

Hearn，Chester G.*Tracks in the Sea*：*Matthew Fontaine Maury and the Mapping of the Oceans*. International Marine/McGraw-Hill，June 2002.

Helland，Pat.“If You Have Too Much Data then “Good Enough' Is

Good Enough."*Communications of the ACM*，June 2011，p.o40 et seq.

Hilbert，Martin，and Priscilla López."The World's Technological Capacity to Store，Communicate，and Compute Information."Science 1（April 2011），pp.60-65.

"How to Measure the World's Technological Capacity to Communicate，Store and Compute Information?"*International Journal of Communication*（2012），pp.1042-55（ijoc.org/ojs/index.php/ijoc/article/viewFile/1562/742）.

Holson，Laura M."Putting a Bolder Face on Google."*New York Times*，March 1，2009，p.BU 1（http：//www.nytimes.com/2009/03/01/business/01marissa.html）.

Hopkins，Brian，and Boris Evelson."Expand Your Digital Horizon with Big Data."Forrester，September 30，2011.

Hotz，Robert Lee."The Really Smart Phone."*Wall Street Journal*，April 22，2011（http：//online.wsj.com/article/SB10001424052748704547604576263261679848814.html）.

Hutchins，John."The First Public Demonstration of Machine Translation：The Georgetown-IBM System，7th January 1954."November 2005（http：//www.hutchinsweb.me.uk/GU-IBM-2005.pdf）.

Inglehart，R.，and H.oD.Klingemann.*Genes*，*Culture and Happiness*.MIT Press，2000.

Isaacson，Walter.*Steve Jobs*.2011.

Kahneman，Daniel.*Thinking*，*Fast and Slow*.Farrar，Straus and Giroux，2011.

Kaplan，Robert S.，and David P.Norton.*Strategy Maps*：*Converting Intangible Assets into Tangible Outcomes*.Harvard Business Review Press，2004.

Karnitschnig，Matthew，and Mylene Mangalindan.“AOL Fires Technology Chief After Web-Search Data Scandal.”*Wall Street Journal*，August 21，2006.

Keefe，Patrick Radden.“Can Network Theory Thwart Terrorists?”*New York Times*，March 12，2006（http：//www.nytimes.com/2006/03/12/magazine/312wwln_essay.html）.

Kinnard，Douglas.*The War Managers*.University Press of New England，1977.

Kirwan，Peter.“This Car Drives Itself.”*Wired UK*，January 2012（http：//www.wired.co.uk/magazine/archive/2012/01/features/this-car-drives-itself）.

Kliff，Sarah.“A Database That Could Revolutionize Health Care.”*Washington Post*，May 21，2012.

Kruskal，William，and Frederick Mosteller.“Representative Sampling，IV：The History of the Concept in Statistics，1895-1939.”*International Statistical Review* 48（1980），pp.169-195.

Laney，Doug.“To Facebook You're Worth $80.95.”*Wall Street Journal*，May 3，2012（http：//blogs.wsj.com/cio/2012/05/03/to-Facebook-youre-worth-80-95/）.

Latour，Bruno，et al.*The Pasteurization of France*.Harvard University Press，1993.

Levitt，Steven D.，and Stephen J.Dubner.*Freakonomics*：*A Rogue Economist Explores the Hidden Side of Everything*.William Morrow，2009.

Levy，Steven.*In the Plex*.Simon and Schuster，2011.

Lewis，Charles Lee.Matthew Fontaine Maury：*The Pathfinder of the Seas*.U.S.Naval Institute，1927.

Lohr，Steve.“Can Apple Find More Hits Without Its Tastemaker?”*New*

York Times，January 19，2011，p.B1（http：//www.nytimes.com/2011/01/19/technology/companies/19innovate.html）.

Lowrey，Annie.“Economists' Programs Are Beating U.S.at Tracking Inflation.”*Washington Post*，December 25，2010（http：//www.washingtonpost.com/wp-dyn/content/article/2010/12/25/AR2010122502600.html）.

Macrakis，Kristie.*Seduced by Secrets*：*Inside the Stasi's Spy-Tech World*.Cambridge University Press，2008.

Manyika，James，et al.“Big Data：The Next Frontier for Innovation，Competition，and Productivity.”McKinsey Global Institute，May 2011（http：//www.mckinsey.com/insights/mgi/research/technology_and_in

Marcus，James.*Amazonia*：*Five Years at the Epicenter of the Dot.Com Juggernaut*.The New Press，2004.

Margolis，Joel M.“When Smart Grids Grow Smart Enough to Solve Crimes.”Neustar，March 18，2010（http：//energy.gov/sites/prod/files/gcprod/documents/Neustar_Comments_DataExhibitA.pdf）.

Maury，Matthew Fontaine.*The Physical Geography of the Sea*.Harper，1855.

Mayer-Schönberger，Viktor.“Beyond Privacy，Beyond Rights：Towards a “Systems”Theory of Information Governance.”98 California Law Review 1853（2010）.

Viktor.Delete：*The Virtue of Forgetting in the Digital Age*.Princeton University Press，2nd ed.，2011.

McGregor，Carolyn，Christina Catley，Andrew James，and James Padbury.“Next Generation Neonatal Health Informatics with Artemis.”In European Federation for Medical Informatics，*User Centred Networked Health Care*，ed.A.Moen et al.（IOS Press，2011），p.117 et seq.

McNamara，Robert S.，with Brian VanDeMark.*In Retrospect*：*The Tragedy and Lessons of Vietnam*.Random House，1995.

Mehta，Abhishek.“Big Data：Powering the Next Industrial Revolution.”Tableau Software White Paper，2011.

Michel，Jean-Baptiste，et al.“Quantitative Analysis of Culture Using Millions of Digitized Books.”*Science* 331（January 14，2011），pp.176-182（http：//www.sciencemag.org/content/331/6014/176.abstract）.

Miller，Claire Cain.“U.S.Clears Google Acquisition of Travel Software.”*New York Times*，April 8，2011（http：//www.nytimes.com/2011/04/09/technology/09google.html?_r=0）.

Mills，Howard.“Analytics：Turning Data into Dollars.”*Forward Focus*，December 2011（http：//www.deloitte.com/assets/Dcom-UnitedStates/Local%20Assets/Documents/FSI/US_FSI_Forward%20Focus_Analytics_Turning%20data%20into%20dollars_120711.pdf）.

Mindell，David A.*Digital Apollo*：*Human and Machine in Spaceflight*.MIT Press，2008.

Minkel，J.R.“The U.S.Census Bureau Gave Up Names of Japanese-Americans in WW II.”*Scientific American*，March 30，2007（http：//www.scientificamerican.com/article.cfm?id=confirmed-the-us-census-b）.

Murray，Alexander.*Reason and Society in the Middle Ages*.Oxford University Press，1978.

Nalimov，E.V.，G.McC.Haworth，and E.A.Heinz.“Space-Efficient Indexing of Chess Endgame Tables.”*ICGA Journal* 23，no.3（2000），pp.148-162.

Narayanan，Arvind，and Vitaly Shmatikov.“How to Break the Anonymity of the Netflix Prize Dataset.”October 18，2006，arXiv：cs/0610105 [cs.CR]（http：//arxiv.org/abs/cs/0610105）.

"Robust De-Anonymization of Large Sparse Datasets."*Proceedings of the 2008 IEEE Symposium on Security and Privacy*，p.111（http：//www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf）.

Nazareth，Rita，and Julia Leite."Stock Trading in U.S.Falls to Lowest Level Since 2008."Bloomberg，August 13，2012（http：//www.bloomberg.com/news/2012-08-13/stock-trading-in-u-s-hits-lowest-level-since-2008-as-vix-falls.html）.

Negroponte，Nicholas.*Being Digital*.Alfred Knopf，1995.

Neyman，Jerzy."On the Two Different Aspects of the Representative Method：The Method of Stratified Sampling and the Method of Purposive Selection."*Journal of the Royal Statistical Society* 97，no.4（1934），pp.558-625.

Ohm，Paul."Broken Promises of Privacy：Responding to the Surprising Failure of Anonymization."57 *UCLA Law Review* 1701（2010）.

Onnela，J.P.，et al."Structure and Tie Strengths in Mobile Communication Networks."*Proceedings of the National Academy of Sciences of the United States of America*（PNAS）104（May 2007），pp.7332-36（http：//nd.edu/~dddas/Papers/PNAS0610245104v1.pdf）.

Palfrey，John，and Urs Gasser.*Interop*：*The Promise and Perils of Highly Interconnected Systems*.Basic Books，2012.

Pearl，Judea.*Models*，*Reasoning and Inference*.Cambridge University Press，2009.

Pollack，Andrew."DNA Sequencing Caught in the Data Deluge."*New York Times*，November 30，2011（http：//www.nytimes.com/2011/12/01/business/dna-sequencing-caught-in-deluge-of-data.html?pagewanted=all）.

President's Council of Advisors on Science and Technology."Report to the President and Congress Designing a Digital Future：Federally Funded Research and Development in Networking and Information

Technology."December 2010（http：//www.whitehouse.gov/sites/default/files/microsites/ost p/pcast-nitrd-report-2010.pdf）.

Priest，Dana and William Arkin."A Hidden World，Growing Beyond Control."*Washington Post*，July 19，2010（http：//projects.washingtonpost.com/top-secret-america/articles/a-hidden-world-growing-beyond-control/print/）.

Query，Tim."Grade Inflation and the Good-Student Discount."Contingencies Magazine，American Academy of Actuaries，May-June 2007（http：//www.contingencies.org/mayjun07/tradecraft.pdf）.

Quinn，Elias Leake."Smart Metering and Privacy：Existing Law and Competing Policies；A Report for the Colorado Public Utility Commission."Spring 2009（http：//www.w4ar.com/Danger_of_Smart_Meters_Colorado_Report.pdf）.

Reshef，David，et al."Detecting Novel Associations in Large Data Sets."*Science*（2011），pp.1518-24.

Rosenthal，Jonathan."Banking Special Report."*The Economist*，May 19，2012，pp.7-8.

Rosenzweig，Phil."Robert S.McNamara and the Evolution of Modern Management."*Harvard Business Review*，December 2010，pp.87-93（http：//hbr.org/2010/12/robert-s-mcnamara-and-the-evolution-of-modern-management/ar/pr）.

Rudin，Cynthia，et al."21st-Century Data Miners Meet 19th-Century Electrical Cables."*Computer*，June 2011，pp.o103-105.

"Machine Learning for the New York City Power Grid."*IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.2（2012），pp.o328-345（http：//hdl.handle.net/1721.1/68634）.

Rys，Michael."Scalable SQL."*Communications of the ACM*，June

2011，48，pp.o48-53.

Salathé，Marcel，and Shashank Khandelwal.“Assessing Vaccination Sentiments with Online Social Media：Implications for Infectious Disease Dynamics and Control.”*PloS Computational Biology* 7，no.10（October 2011）.

Savage，Mike，and Roger Burrows.“The Coming Crisis of Empirical Sociology.”*Sociology* 41（2007），pp.885-899.

Schlie，Erik，J.rg Rheinboldt，and Niko Waesche.*Simply Seven*：*Seven Ways to Create a Sustainable Internet Business*.Palgrave Macmillan，2011.

Scanlon，Jessie.“Luis von Ahn：The Pioneer of ‘Human Computation’.”*Businessweek*，November 3，2008（http：//www.businessweek.com/stories/2008-11-03/luis-von-ahn-the-pioneer-of-human-computation-businessweek-business-news-stock-market-and-financial-advice）.

Scism，Leslie，and Mark Maremont.“Inside Deloitte's Life-Insurance Assessment Technology.”*Wall Street Journal*，November 19，2010（http：//online.wsj.com/article/SB10001424052748704104104575622531084755588.html）.

“Insurers Test Data Profiles to Identify Risky Clients.”*Wall Street Journal*，November 19，2010（http：//online.wsj.com/article/SB10001424052748704648604575620750998072986.html）.

Scott，James.*Seeing like a State*：*How Certain Schemes to Improve the Human Condition Have Failed*.Yale University Press，1998.

Seltzer，William，and Margo Anderson.“The Dark Side of Numbers：The Role of Population Data Systems in Human Rights Abuses.”*Social Research* 68（2001）pp.481-513.

Silver，Nate.*The Signal and the Noise*：*Why So Many Predictions*

Fail-But Some Don't.Penguin，2012.

Singel，Ryan.“Netflix Spilled Your Brokeback Mountain Secret，Lawsuit Claims.”*Wired*, December 17，2009（http：//www.wired.com/threatlevel/2009/12/netflix-privacy-lawsuit/）.

Smith，Adam.*The Wealth of Nations*（1776）.Reprinted Bantam Classics，2003.A free electronic version is available（http：//www2.hn.psu.edu/faculty/jmanis/adam-smith/Wealth-Nations.pdf）.

Solove，Daniel J.*The Digital Person*：*Technology and Privacy in the Information Age*.NYU Press，2004.

Surowiecki，James.“A Billion Prices Now.”*New Yorker*，May 30，2011（http：//www.newyorker.com/talk/financial/2011/05/30/110530ta_talk_surowiecki）.

Taleb，Nassim Nicholas.*Fooled by Randomness*：*The Hidden Role of Chance in Life and in the Markets*.Random House，2008.

The Black Swan：*The Impact of the Highly Improbable*.2nd ed.，Random House，2010.

Thompson，Clive.“For Certain Tasks，the Cortex Still Beats the CPU.”*Wired*，June 25，2007（http：//www.wired.com/techbiz/it/magazine/15-07/ff_humancomp?currentPage=all）.

Thurm，Scott.“Next Frontier in Credit Scores：Predicting Personal Behavior.”*Wall Street Journal*，October 27，2011（http：//online.wsj.com/article/SB10001424052970203687504576655182086300912.html）.

Tsotsis，Alexia.“Twitter Is at 250 Million Tweets per Day，iOS 5 Integration Made Signups Increase 3x.”*TechCrunch*，October 17，2011（http：//techcrunch.com/2011/10/17/Twitter-is-at-250-million-tweets-

per-day/）.

Valery，Nick.“Tech.View：Cars and Software Bugs.”*The Economist*，May 16，2010（http：//www.economist.com/blogs/babbage/2010/05/techview_cars_and_software_bugs）.

Vlahos，James.“The Department Of Pre-Crime.”*Scientific American* 306（January 2012），pp.62-67.

Von Baeyer，Hans Christian.*Information*：*The New Language of Science*.Harvard University Press，2004.

von Ahn，Luis，et al.“reCAPTCHA：Human-Based Character Recognition via Web Security Measures.”*Science* 321（September 12，2008），pp.1465-68（http：//www.sciencemag.org/content/321/5895/1465.abstract）.

Watts，Duncan.*Everything Is Obvious Once You Know the Answer*：*How Common Sense Fails Us*.Atlantic，2011.

Weinberger，David.*Everything Is Miscellaneous*：*The Power of the New Digital Disorder*.Times，2007.

Weinberger，Sharon.“Intent to Deceive.”*Nature* 465（May 2010），pp.412-415（http：//www.nature.com/news/2010/100526/full/465412a.html）.

“Terrorist ‘Pre-crime’ Detector Field Tested in United States.”*Nature*，May 27，2011（http：//www.nature.com/news/2011/110527/full/news.2011.323.html）.

Whitehouse，David.“UK Science Shows Cave Art Developed Early.”BBC News Online，October 3，2001（http：//news.bbc.co.uk/1/hi/sci/tech/1577421.stm）.

Wigner，Eugene.“The Unreasonable Effectiveness of Mathematics in

the Natural Sciences."*Communications on Pure and Applied Mathematics* 13，no.1（1960），pp.o1-14.

Wilks，Yorick.*Machine Translation*：*Its Scope and Limits*.Springer，2009.

Wingfield，Nick."Virtual Products，Real Profits：Players Spend on Zynga's Games，but Quality Turns Some Off."*Wall Street Journal*，September 9，2011（http：//online.wsj.com/article/SB10001424053111904823804576502442835413446.html）.